J. R. Statist. Soc. B (1998)
60, *Part* 1, *pp.* 3–21

Robust models in probability sampling

D. Firth†

University of Oxford, UK

and K. E. Bennett

Medical Research Council Institute of Hearing Research, Nottingham, UK

[*Read before* The Royal Statistical Society *at a meeting on the 'Design and analysis of complex sample surveys' organized by the* Research Section *on Wednesday, May 14th, 1997,* Dr D. Holt *in the Chair*]

Summary. In the estimation of a population mean or total from a random sample, certain methods based on linear models are known to be automatically design consistent, regardless of how well the underlying model describes the population. A sufficient condition is identified for this type of robustness to model failure; the condition, which we call 'internal bias calibration', relates to the combination of a model and the method used to fit it. Included among the internally bias-calibrated models, in addition to the aforementioned linear models, are certain canonical link generalized linear models and nonparametric regressions constructed from them by a particular style of local likelihood fitting. Other models can often be made robust by using a suboptimal fitting method. Thus the class of model-based, but design consistent, analyses is enlarged to include more realistic models for certain types of survey variable such as binary indicators and counts. Particular applications discussed are the estimation of the size of a population subdomain, as arises in tax auditing for example, and the estimation of a bootstrap tail probability.

Keywords: Auditing; Bias calibration; Bootstrap acceleration; Control variate; Finite population; Generalized linear model; Importance sampling; Instrumental variable; Local likelihood; Logistic regression; Smoothing; Spline; Stratification; Survey sampling

1. Introduction

The use of regression models to derive estimators of population quantities such as means and totals is well established. Examples may be found in standard texts such as Cochran (1977), and a fairly recent survey is Särndal *et al.* (1992). It is well known that estimators based on certain linear model specifications enjoy a robustness property, namely that they are design consistent for the target quantity even under gross failure of the model on which they are based. A familiar example is the ratio estimator

$$\hat{T} = N^{-1} \sum_{i=1}^{N} x_i \Big(\sum_s y_j \Big/ \sum_s x_j \Big)$$

of a finite population mean

$$T = N^{-1} \sum_{i=1}^{N} y_i,$$

when s is a simple random sample; here $\hat{T}$ may be derived as the best linear unbiased predictor of T under the linear model specification $E(y_i) = x_i\beta$, $\mathrm{var}(y_i) = \sigma^2 x_i$. In repeated

†*Address for correspondence*: Nuffield College, New Road, Oxford, OX1 1NF, UK.
E-mail: david.firth@nuf.ox.ac.uk

 1369–7412/98/60003

sampling, when the sample size is large, $\hat{T}$ is approximately unbiased for T regardless of whether the corresponding linear model approximately represents the population regression.

The class of linear-model-based estimators that are asymptotically design unbiased, or design consistent, and in that sense robust to model failure, has been explored by several researchers, e.g. Little (1983), Robinson and Särndal (1983), Wright (1983) and Särndal and Wright (1984). The technical distinction between asymptotic design unbiasedness and design consistency will be unimportant in what follows. Both are defined with reference to an appropriate asymptotic framework that allows n and N to increase together under some fairly mild restrictions, and, provided that $\mathrm{var}(\hat{T})/T^2 \to 0$ as $n,\ N \to \infty$, asymptotic design unbiasedness implies design consistency. Various specific formulations of the asymptotic framework are possible, with minor differences which also are not crucial to the development below. The essential feature of estimators with the desired properties is that their bias, in repeated sampling, is an order of magnitude smaller than their standard deviation if the sample size is sufficiently large. The present work identifies some types of non-linear regression model, including certain generalized linear and nonparametric regressions, which have this robustness property for estimating a population mean or total. The class of 'robust models' is thereby extended to include more realistic representations of populations involving, for example, a binary indicator or a count as the main survey variable, and further to accommodate complex patterns of non-linearity that are not adequately described by linear or even generalized linear models.

Suppose that the quantity of interest is of the general form

$$T = \sum_{i=1}^{N} a_i\, y_i,$$

where i indexes population units, N is the population size and $(y_i,\ a_i)$ are values associated with each unit; before sampling, a_i is known and y_i unknown for each i. Familiar examples are $a_i \equiv 1$ and $a_i \equiv 1/N$, where respectively the population total and the population mean are the target quantities. The work reported here was motivated initially by a tax auditing application, discussed further in Section 3.1 later, in which each of N transactions of a company under audit is either taxable ($y_i = 1$) or not ($y_i = 0$), the size of the ith transaction being the known amount of money a_i, and $T = \Sigma_{i=1}^{N} a_i y_i$ being the total amount taxable. The same estimation problem arises also in other application contexts; for example in market research a_i may be the turnover of the ith in a population of farms, and $y_i = 1$ if the farm takes a particular periodical, so that $\Sigma_1^N a_i y_i$ is the total turnover of subscribers.

To estimate T, a random sample s of size n is drawn from the units $\{1, \ldots, N\}$. Most often in practice s is drawn without replacement, but much of the following discussion may be straightforwardly adapted also to with-replacement sampling. Write $p(s)$ for the probability of selecting each particular sample s, and denote the first-order inclusion probabilities by $\pi_i = P(i \in s)$ $(i = 1, \ldots, N)$, assumed to be all non-zero. In what follows, the sampling plan is taken as fixed, the focus being on models and associated estimators, i.e. design issues are regarded as outside the scope of the present paper.

Operationally, a model for the population values $\{y_i\colon\ i = 1, \ldots, N\}$ provides a set $\mathcal{M} = \{\hat{y}_i\colon\ i = 1, \ldots, N\}$ of predicted values. Two alternative estimates of T derived from $\mathcal{M}$ are then the 'projective' form

$$\hat{T}_{\mathrm{pro}} = \sum_{i=1}^{N} a_i \hat{y}_i \tag{1}$$

and the 'predictive' form

$$\hat{T}_{\text{pre}} = \sum_{i \in s} a_i y_i + \sum_{i \notin s} a_i \hat{y}_i, \tag{2}$$

in terminology similar to that of Särndal and Wright (1984). If the finite population values $y_1, \ldots, y_N$ are themselves viewed as a random sample from a 'superpopulation' described by the model (e.g. Royall (1970)), then estimation of T is a prediction problem, and $\hat{T}_{\text{pre}}$ is preferred to $\hat{T}_{\text{pro}}$ in situations where the two forms differ. Some of the most commonly used linear regression approaches involve estimation procedures which ensure that

$$\sum_s a_i \hat{y}_i = \sum_s a_i y_i$$

for all samples s, so that $\hat{T}_{\text{pre}} \equiv \hat{T}_{\text{pro}}$; the above ratio estimator of $T = N^{-1} \Sigma\, y_i$, with

$$\hat{y}_i = x_i \left(\sum_s y_j \Big/ \sum_s x_j \right),$$

is a particular instance. Under simple random sampling and other equal π_i designs, this 'predictive–projective equivalence' (PPE) property has a certain intuitive appeal; moreover, as will be shown in Section 2.1, under equal probability sampling PPE is sufficient for design consistency of $\hat{T}$, where $\hat{T}$ here denotes either of the equivalent forms $\hat{T}_{\text{pro}}$ and $\hat{T}_{\text{pre}}$. More generally, with unequal probability designs, the corresponding sufficient condition is essentially a weighted version of PPE, with weights determined by the reciprocal inclusion probabilities π_i^{-1}.

The paper is organized as follows. In Section 2 a sufficient condition is identified under which a model-based estimator as in equation (1) or (2) is design consistent, and implications of that condition are explored in the broad classes of linear, generalized linear, stratified and nonparametrically smooth 'local' regression models. Section 3 discusses two particular applications: estimation of the size of a population subdomain, as in the tax auditing or market research contexts outlined above, and improved bootstrap tail probability estimation by logistic regression on a control variate. Some brief concluding remarks are collected in Section 4.

It should be noted at the outset that the title of this paper is a slight over-simplification, even if the rather special interpretation of 'robust' is accepted. As has been hinted at above, and as will be evident in Section 2, the condition for design consistency is a condition on the *combination* of a model and an associated fitting procedure. Thus, for example, in the case of the ratio estimator above, it is the pairing of

(a) the linear model $E(y_i) = x_i\beta$, $\text{var}(y_i) = \sigma^2 x_i$, and
(b) the calculation of predicted values by weighted least squares estimation of β from the sample

which results in the PPE property and hence, under simple random sampling, design consistency. In general, especially appealing model–method combinations are of course those in which the method is optimal for the model, as with the optimally weighted least squares estimation used in the ratio estimator example above. In the wider context of generalized linear models, this consideration will be found to provide a special role for models with canonical link; for example in the case of binary y_i the maximum likelihood fit of a suitably specified logistic regression can be used directly to yield a design consistent estimator of T, but the same is not true of, for example, probit or complementary log–log-regressions. A

model whose optimal fitting method fails to yield a design consistent $\hat{T}$ can often be made robust by using an appropriately modified, suboptimal fitting procedure; some examples of this are discussed briefly in Sections 2.2.2 and 2.2.3.

2. Robust model-based estimators

2.1. Internally bias-calibrated models

In typical applications $\hat{y}_i$ is derived from a regression formula $\hat{y}_i = \hat{\mu}(\mathbf{x}_i)$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is a vector of p known covariate values associated with each unit ($i = 1, \ldots, N$), and $\hat{\mu}(\cdot)$ is a regression function, either known in advance or calculated from the sample data $\{(y_i, \mathbf{x}_i): \ i \in s\}$. If $\hat{\mu}(\cdot)$ does not depend on s, the estimator

$$\hat{T}_{\text{diff}} = \sum_{i=1}^{N} a_i \hat{\mu}(\mathbf{x}_i) + \sum_{i \in s} \frac{a_i}{\pi_i} \{y_i - \hat{\mu}(\mathbf{x}_i)\} \tag{3}$$

is known as a *difference estimator*, is easily shown to be exactly design unbiased for T and under weak limiting assumptions is design consistent as the sample size increases (e.g. Cassel *et al.* (1976) and Särndal *et al.* (1992)). A simple, well-known example is the 'Horvitz–Thompson' unbiased estimator

$$\hat{T}_{\text{diff}} = \sum_{s} \pi_i^{-1} a_i y_i,$$

which results from taking $\hat{\mu}(\mathbf{x}_i) \equiv 0$.

If $\hat{\mu}(\cdot)$ is estimated from the sample, $\hat{T}_{\text{diff}}$ as in equation (3) is no longer exactly unbiased in repeated sampling, but under further weak assumptions is asymptotically unbiased and, as a consequence, remains design consistent for T. In the particular case where $\hat{\mu}(\mathbf{x}_i)$ is a linear regression with coefficients estimated from the sample by ordinary or weighted least squares, $\hat{T}_{\text{diff}}$ is known as a *generalized regression estimator* (e.g. Cassel *et al.* (1976), Särndal (1980) and Särndal *et al.* (1992)). In essence, provided that $\hat{\mu}(\mathbf{x}_i)$ converges in probability, at an appropriate rate, to some fixed function $\mu(\mathbf{x}_i)$ as the sample size increases, $\hat{T}_{\text{diff}}$ in equation (3) behaves like a difference estimator in sufficiently large samples, and in particular is design consistent; and this argument is easily seen to apply not only to linear models but also in general. The estimator $\hat{T}_{\text{diff}}$ may be viewed as a 'bias-calibrated' version of the simple model projection estimator $\hat{T}_{\text{pro}} = \Sigma_1^N a_i \hat{\mu}(\mathbf{x}_i)$: the extra term $\Sigma_s \pi_i^{-1} a_i \{y_i - \hat{\mu}(\mathbf{x}_i)\}$ estimates the population sum $\Sigma_1^N a_i \{y_i - \hat{\mu}(\mathbf{x}_i)\}$ and thus removes, approximately, the bias of $\hat{T}_{\text{pro}}$ in repeated sampling. Alternatively, $\hat{T}_{\text{diff}}$ may be expressed in terms of the predictive estimator, $\hat{T}_{\text{pre}} = \hat{T}_{\text{pro}} + \Sigma_s a_i \{y_i - \hat{\mu}(\mathbf{x}_i)\}$, as

$$\hat{T}_{\text{diff}} = \hat{T}_{\text{pre}} + \sum_{i \in s} a_i (\pi_i^{-1} - 1)\{y_i - \hat{\mu}(\mathbf{x}_i)\},$$

in which the second term again removes the estimated bias. The terminology 'bias calibrated' is borrowed from Chambers *et al.* (1993), where a similar device is used to reduce bias under the model due to smoothing.

Consider now a general model-based estimator $\hat{T}$ in one of the two forms $\hat{T}_{\text{pre}}$ or $\hat{T}_{\text{pro}}$ of equations (1) and (2). For each population unit i, define $q_i = \pi_i^{-1}$ if $\hat{T}$ is of the form $\hat{T}_{\text{pro}}$, or $q_i = \pi_i^{-1} - 1$ if $\hat{T}$ is of the form $\hat{T}_{\text{pre}}$. Then, if

$$\sum_{i \in s} q_i a_i \{y_i - \hat{\mu}(\mathbf{x}_i)\} = 0 \tag{4}$$

for all possible samples s, the combination of model and associated fitting procedure, which together determine the mapping $s \mapsto \hat{\mu}(\cdot)$, will be called *internally bias calibrated* (IBC) for T through $\hat{T}$. The following observations are immediate from the discussion above:

(a) if the model and its associated fitting procedure are IBC for T through $\hat{T}$, then $\hat{T}$ is design consistent for T;
(b) under equal probability sampling schemes with $\pi_i = n/N$ for all units, e.g. simple random sampling or stratified random sampling with proportional allocation, the IBC and PPE properties coincide.

In the rest of this paper, terms such as 'IBC model' etc. will often be used as a convenient shorthand to refer to a pairing of a model and fitting procedure which is IBC for a specified target quantity T through one or both of $\hat{T}_{\text{pre}}$ and $\hat{T}_{\text{pro}}$, with appropriate qualification being provided when such usage is ambiguous.

The IBC condition (4), although sufficient, is by no means necessary for design consistency; the informal argument above indicates that if condition (4) holds in expectation under repeated sampling, or indeed if

$$E\left[\sum_{i \in s} q_i a_i y_i\right] = E\left[\sum_{i \in s} q_i a_i \hat{\mu}(\mathbf{x}_i)\right] + b_n$$

where b_n is sufficiently small as n increases, the same result holds, i.e. $\hat{T}$ is approximately design unbiased, and hence under standard conditions design consistent, for T. In the following, however, the focus is on models which satisfy condition (4) exactly for all samples, since — as will be shown — this seems a sufficiently rich class of models for many practical purposes.

Models with the IBC property, compared with those requiring 'external' bias calibration as in $\hat{T}_{\text{diff}}$, are attractive from several viewpoints. From a model-based standpoint, Little (1983) argued strongly that it is more 'principled' to build models which automatically yield design consistent estimators than to make a post-modelling adjustment such as the bias calibration in equation (3). From the opposite viewpoint of purely design-based inference, wherein the model has no status other than as a device used to generate an estimator, it may be argued that estimators of the form $\hat{T}_{\text{pre}}$ or $\hat{T}_{\text{pro}}$ are appealing mainly for their 'cosmetic' simplicity (Särndal and Wright, 1984), relative to an externally bias-calibrated form as in equation (3); there may be advantages also in the simplicity with which the estimator's design variance can be estimated, especially in the case of $\hat{T}_{\text{pro}}$ derived from a non-linear parametric model, since general results such as those of Binder (1983) apply whenever the projected regression formula is estimated by solving score-like equations. A third view is provided by the theoretical framework of Godambe and Thompson (1986), in which the IBC condition emerges as an implicit requirement of the optimum estimating functions approach; see Section 4 for brief further discussion.

The remaining parts of this section explore the construction of IBC models of various kinds. For linear models, design consistency has been extensively studied, and the property which we have called IBC is certainly not new; see, for example Brewer (1979), Särndal (1980), Little (1983), Wright (1983), Mantel (1991) and especially Isaki and Fuller (1982), Särndal and Wright (1984), Brewer *et al.* (1988) and Brewer (1995). The main aim of this paper is to extend the same ideas to non-linear regression models. The known results for

linear models are encompassed in Section 2.2.1 below, where a linear model is viewed as a particular instance of a canonical link generalized linear model.

2.2. Generalized linear models

2.2.1. Canonical link models

For calculation of the coefficients $\hat{\beta}$ in the generalized linear model

$$\hat{\mu}(\mathbf{x}_i) = g^{-1}(\hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}), \tag{5}$$

consider the system of weighted estimating equations

$$u_r = \sum_{i \in s} w_{ir} \frac{y_i - \hat{\mu}(\mathbf{x}_i)}{V\{\hat{\mu}(\mathbf{x}_i)\}} \frac{x_{ir}}{g'\{\hat{\mu}(\mathbf{x}_i)\}} = 0 \qquad (r = 1, \ldots, p). \tag{6}$$

Here $g(\cdot)$ and $V(\cdot)$ are respectively the link and variance functions for the generalized linear model and the $\{w_{ir}\}$ are fixed constants: if $w_{ir} \equiv w_i$, the $\{w_i\}$ are straightforward case weights, and more generally w_{ir} may be viewed as the relative weight of the ith unit in the rth equation. When $w_{ir} \equiv 1$, equations (6) are the standard maximum likelihood equations for the exponential family generalized linear model with $\hat{\mu}(\mathbf{x}_i)$ as in equation (5) and $\mathrm{var}(y_i)$ proportional to $V(\hat{\mu})$, or alternatively they are best linear ('quasi-likelihood') estimating equations for the same mean–variance specification; see McCullagh and Nelder (1989) for a full account. Case weights w_i may be used, for example, to accommodate a known pattern of non-proportionality between $\mathrm{var}(y_i)$ and $V(\mu_i)$ or may be chosen to depend in some way on the inclusion probabilities π_i. For estimating the general target quantity $T = \Sigma_1^N a_i y_i$, it follows from equation (6) that the model has the IBC property if

(a) $V(\mu) \propto 1/g'(\mu)$ and
(b) $q_i a_i$ is a linear combination of $(w_{i1}x_{i1}, \ldots, w_{ip}x_{ip})$, say

$$q_i a_i = \lambda_1 w_{i1} x_{i1} + \ldots + \lambda_p w_{ip} x_{ip} \qquad (i = 1, \ldots, N).$$

For, if (a) and (b) are satisfied,

$$\sum_{i \in s} q_i a_i \{y_i - \hat{\mu}(\mathbf{x}_i)\} = \sum_{r=1}^{p} \lambda_r u_r = 0,$$

as required.

Condition (a) above is satisfied whenever the link function $g(\cdot)$ is canonical, relative to variance function $V(\cdot)$; common examples are the linear model with variance free of the mean ($g(\mu) = \mu$, $V(\mu) = 1$), log-linear models with variance proportional to the mean ($g(\mu) = \log(\mu)$, $V(\mu) = \mu$) and logit models for binary data ($g(\mu) = \log\{\mu/(1-\mu)\}$, $V(\mu) = \mu(1-\mu)$).

Condition (b) may be satisfied by an appropriate choice of covariates x_{ir} and weights w_{ir}, the considerations involved in this choice being essentially the same as in linear models.

As a simple, but familiar and important, example, consider the estimation of the population mean $N^{-1}\Sigma_1^N y_i$ under an equal probability plan such as simple random sampling. Here a_i and q_i are both constant, and we see that any canonical link generalized linear model containing an intercept term ($x_{i1} \equiv 1$, say), fitted by ordinary, unweighted maximum likelihood, has the IBC property, which in this case coincides with the PPE property introduced in Section 1. For concreteness, suppose that a single auxiliary variable x is available. Familiar estimators of the population mean are then the linear regression estimator based on the model

$$\hat{\mu}(x_i) = \hat{\beta}_1 + \hat{\beta}_2 x_i, \qquad V(\hat{\mu}) = 1, \tag{7}$$

and the ratio estimator as in Section 1, which is based on

$$\hat{\mu}(x_i) = \hat{\beta} x_i, \qquad V(\hat{\mu}) = \hat{\mu}. \tag{8}$$

The first of these is plainly a canonical link model with an intercept term, whereas the 'ratio' model (8) may be re-expressed as

$$\log\{\hat{\mu}(x_i)\} = \log(\hat{\beta}) + \log(x_i), \qquad V(\hat{\mu}) = \hat{\mu}, \tag{9}$$

i.e. as a canonical link model in which $\log(\hat{\beta})$ is the intercept term. Of course, it may be that neither model (7) nor model (8) is a good description of the population scatter of *y versus x*. One possibility, even if the assumed mean–variance relationship is adequate, is non-linearity of $\hat{\mu}(x)$ in x. The linear model (7) can be elaborated directly, by adding further terms such as polynomials or regression splines to the linear predictor, to accommodate non-linearity while maintaining the IBC property. If the ratio model (8) is to be similarly elaborated, while maintaining both the IBC property and regression through the origin, $\hat{\mu}(0) = 0$, an appropriate route would be the addition of extra terms to the linear predictor of the log-linear representation (9); the direct addition of further terms to $\hat{\beta} x_i$ in model (8) would violate the IBC property, and thus in general would yield a design inconsistent estimator. A further possibility is that models (7) and (8) are both unrealistic on account of the assumed mean–variance relationships. For example, if y_i is binary the appropriate mean–variance relationship is necessarily $V(\mu) = \mu(1 - \mu)$, and a more appropriate model would then be a logistic regression; any logistic regression model containing an intercept term is IBC for $N^{-1} \Sigma_1^N y_i$. More generally, when $T = \Sigma_1^N a_i y_i$ with the $\{a_i\}$ unequal, the maximum likelihood fit of any canonical link model with a_i in the span of $x_{i1}, \ldots, x_{ip}$ is IBC under simple random sampling: for an application involving binary y_i, see Section 3.1.

In general, there is an infinite number of ways to achieve condition (b) by weighting and/or the inclusion of suitable extra terms in the linear predictor. To illustrate, consider the use of $\hat{T}_{\text{pre}}$ based on the simple linear model

$$\hat{\mu}(x_i) = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{10}$$

to estimate $T = N^{-1} \Sigma_1^N y_i$ under an unequal probability sampling plan. If model (10) is fitted by ordinary least squares, the IBC condition requires $q_i = \pi_i^{-1} - 1$ to be a linear combination of $\{1, x_i\}$, which will not usually be the case. Three specific prescriptions to achieve IBC in this instance are

(a) fit model (10) using case weights $w_{ir} \equiv w_i = \pi_i^{-1} - 1$ (prescription 1);
(b) fit model (10) using instrumental variable estimation (Brewer *et al.*, 1988; Brewer, 1995) by setting, for example, $w_{i1} = 1$ and $w_{i2} = \pi_i^{-1}/x_i$ so that π_i^{-1} is in effect used as an instrument for x_i in estimating the model (prescription 2);
(c) include π_i^{-1} as an extra covariate, and fit without weights, i.e. use 'weights' $w_{ir} \equiv 1$ (prescription 3).

From a purely model-based viewpoint, prescription 3 seems preferable as it avoids the use of possibly suboptimal weights: optimum estimation under the model requires weights proportional to $1/\text{var}(y_i)$, so $w_{ir} \equiv 1$ is optimal if variation around model (10) is thought to be homoscedastic. For an alternative, design-based perspective on the choice of weights, see Särndal (1980). Note that under model (10) all three prescriptions are suboptimal, in the

case of prescription 3 not because of inefficient weighting but on account of the redundant covariate π_i^{-1}. A possible advantage of the 'redundant' covariate, though, is a contribution to model robustness if the residual variation around model (10) is correlated with π_i^{-1}. The instrumental variable approach exemplified in prescription 2 provides a flexible alternative to pure case weighting and if the $\{w_{ir}\}$ are chosen optimally will often lose little in terms of efficiency; a more detailed assessment is in preparation as a separate paper.

Little (1983) suggested model elaboration in the spirit of prescription 3 above, but involving the inclusion of a separate indicator variable for each distinct value of π_i used in the sampling plan. The set of such indicator variables includes q_i in its span, and so the IBC condition holds for the model thus elaborated, whether T is estimated through $\hat{T}_{\text{pre}}$ or $\hat{T}_{\text{pro}}$. If the number of distinct values of π_i is large — the extreme situation being where each unit has its own unique value of π_i — this approach results in a model with a very large number of 'nuisance' parameters. Little (1983) recognized this problem and proposed a random effects formulation to overcome it; the inclusion of a single extra covariate as in prescription 3 is an alternative means of incorporating parsimoniously the dependence on π_i needed to ensure design consistency.

Finally we note that, for estimation via the projective form $\hat{T}_{\text{pro}}$, the route to the IBC property corresponding to prescription 1 above is to use case weights $w_{ir} \equiv w_i = \pi_i^{-1}$, as in the method of 'pseudomaximum likelihood' for estimation of the model coefficients (e.g. Binder (1983) and Skinner *et al.* (1989), section 3.4). In general, then, the pseudomaximum likelihood fit of a canonical link model with a_i in the span of $\{x_{i1}, \ldots, x_{ip}\}$ is IBC for $T = \Sigma_1^N a_i y_i$ through $\hat{T}_{\text{pro}} = \Sigma_1^N a_i \hat{\mu}(\mathbf{x}_i)$.

2.2.2. *Non-canonical models*

The essence of the previous section is that, because of the cancellation of V and $1/g'$ in the estimating equations (6), canonical link models fitted by maximum likelihood are subject to precisely the same considerations as linear models fitted by least squares, with regard to the use of weights, instrumental variables and/or extra terms in the linear predictor to ensure design consistency. The same can be made to apply also to non-canonical links if the weighted maximum likelihood equations (6) are replaced by

$$u_r^* = \sum_{i \in s} w_{ir}\{y_i - \hat{\mu}(\mathbf{x}_i)\}x_{ir} = 0 \qquad (r = 1, \ldots, p) \tag{11}$$

for the calculation of $\hat{\beta}_1, \ldots, \hat{\beta}_p$; note that equations (11) and (6) are identical in the case of a canonical link. Thus, if g is non-canonical relative to the assumed form of $V(\cdot)$, the IBC property can still be made to hold by using equations (11) to fit the model, combined with an appropriate choice of weights and covariates as discussed earlier. In general the resultant estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are then suboptimal under the model. In some situations the loss of efficiency is small; for example, in the estimation of a probit regression for binary y_i, equations (11) and (6) are rather similar because $V(\mu) = \mu(1 - \mu)$ is approximately proportional to $1/g'(\mu) = \phi\{\Phi^{-1}(\mu)\}$. In other contexts model efficiency may be partially recoverable by choosing case weights $w_{ir} \equiv w_i$ to be such that $V\{\hat{\mu}(\mathbf{x}_i)\}\, g'\{\hat{\mu}(\mathbf{x}_i)\}/w_i$ is more nearly constant than Vg'.

2.2.3. *Ordered multinomial regression*

We note here one further, important class of models in which the maximum likelihood fit does not in general yield a design consistent estimator, but where a suboptimal fitting method may be used to achieve that end.

Suppose that ordered response categories are labelled 1, . . ., K, that $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$ for the ith unit is an indicator vector with elements

$$y_{ik} = \begin{cases} 1 & \text{if unit } i \text{ is in category } k, \\ 0 & \text{otherwise} \end{cases}$$

and that interest is in the vector of population proportions $\mathbf{T} = N^{-1}\Sigma_1^N \mathbf{y}_i$. A commonly used class of models relating $\mathbf{y}_i$ to auxiliary variables $x_{i1}, \ldots, x_{ip}$ is the class of 'cumulative link' regressions of the form

$$\hat{\mu}_k(\mathbf{x}_i) = g^{-1}(\hat{\theta}_k + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}) \qquad (k = 1, \ldots, K-1),$$

where $g(\cdot)$ is a link function such as the logit or probit and $\hat{\mu}_k(\mathbf{x}_i)$ models the regression of the cumulative indicator $y_{i1} + y_{i2} + \ldots + y_{ik} = z_{ik}$, say; see, for example, McCullagh and Nelder (1989), section 5.2, or Agresti (1989), chapter 9. The multinomial likelihood equations for such a model are considerably more complicated than equations (6), and even with the logit link they do not simplify to a form which allows the IBC property to be engineered by a suitable choice of weights and/or extra covariates.

A simple alternative approach to fitting such a model is to treat the $\{z_{ik}: k = 1, \ldots, K-1\}$ as if they were independent binary observations, i.e. in effect to use a system of 'generalized estimating equations' with working correlations all set to 0 (Liang and Zeger, 1986). Design consistent estimation of each of the cumulative proportions $T_1 + \ldots + T_k$, and hence of $\mathbf{T}$, is then achieved directly as in Section 2.2.1 if g is the logit link, or as in Section 2.2.2 otherwise. This incurs a loss of efficiency under the model, since in reality the $\{z_{ik}: k = 1, \ldots, K-1\}$ are of course correlated. Some preliminary investigations in Wolfe (1996) suggest that the loss of efficiency due to using 'independence' estimating equations for this purpose is often small and may thus be a reasonable price to pay for design consistency.

2.3. *Stratified models*

As one approach to model robustness, the population may be divided into strata, or groups, labelled $g = 1, \ldots, G$, and separate models $\hat{\mu}_g(\mathbf{x}_i)$ used locally within each stratum. The corresponding 'global' regression function is then

$$\hat{\mu}(\mathbf{x}_i) = \sum_g \hat{\mu}_g(\mathbf{x}_i) I_{ig},$$

where $\{I_{ig}: g = 1, \ldots, G\}$ are binary indicators of stratum membership. The target quantity $T = \Sigma_1^N a_i y_i$ may be similarly decomposed as $T = \Sigma_g T_g$, where

$$T_g = \sum_1^N a_i y_i I_{ig},$$

and clearly if each of the $\hat{\mu}_g(\cdot)$ is IBC for the corresponding stratum target T_g then $\hat{\mu}(\cdot)$ is IBC for T. If the same strata are used also in the sampling design, in such a way that $p(s)$ implies constant π_i sampling within each stratum, the IBC property is achieved globally by simply ensuring PPE for each of the within-stratum models, irrespective of the rates at which different strata are sampled.

A simple example is the so-called separate ratio estimator for $T = N^{-1}\Sigma_1^N y_i$ in the presence of a scalar covariate x_i; in the case of constant π_i within strata,

$$\hat{T} = N^{-1} \sum_{i=1}^{N} x_i \sum_{g=1}^{G} \hat{\beta}_g I_{ig},$$

with

$$\hat{\beta}_g = \sum_{i \in s} y_i I_{ig} \Big/ \sum_{i \in s} x_i I_{ig}. \tag{12}$$

The implicit model underlying equation (12) is a separate linear regression through the origin within each stratum. Royall and Herson (1973) demonstrated that within-stratum modelling in this fashion provides a degree of robustness to failure of the model underlying the standard, 'global' ratio estimator introduced in Section 1.

2.4. 'Smooth' local models

Local models as in Section 2.3 are appealing on account of their semiparametric flavour and comparative robustness relative to global parametric models, but the potentially large discontinuities in $\hat{\mu}(\mathbf{x})$ at stratum boundaries will be implausible in many contexts and may have an adverse effect on the efficiency with which population quantities are estimated.

A smooth version of the same idea is as follows: for each population unit i, obtain $\hat{\mu}(\mathbf{x}_i)$ from a local likelihood fit (Tibshirani and Hastie, 1987) of some simple model that has the IBC property. Local likelihood fitting typically employs kernel weights based on distances $|\mathbf{x}_i - \mathbf{x}_j|$, but in general kernel-weighted local fits, even of models which themselves have the IBC property, fail to yield a result with that same property. In the following, a general method is given for obtaining a regression function $\hat{\mu}(\mathbf{x}_i)$, by a construction based on local likelihood fitting, which has the required IBC property. For simplicity, it will be assumed that a single auxiliary variable x is available; the approach can easily be extended in principle to multidimensional $\mathbf{x}$, but the amount of computation involved grows rapidly.

Suppose that $\hat{\mu}(x)$, and its associated fitting procedure, together comprise a model which is IBC for $\Sigma_1^N a_i y_i$ through one or both of $\hat{T}_{\text{pre}}$ and $\hat{T}_{\text{pro}}$; the simplest example, for $a_i \equiv 1/N$ and with equal probability sampling, is

$$\hat{\mu}(x) = n^{-1} \sum_{s} y_i,$$

in which dependence on x is null. Let the members of s have (x, y) values $\{(x_{(1)}, y_{(1)}), \ldots, (x_{(n)}, y_{(n)})\}$, arranged so that $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$. Let k be some fixed positive integer: the constant k will determine the effective bandwidth, or alternatively the effective degrees of freedom, of the smoothing method proposed here. Denote by $\hat{\mu}_r(x)$, for $r \in \{2-k, \ldots, n\}$, the fit of the model to the subset $s(r)$ of the sample, where

$$s(r) = \begin{cases} \{(x_{(1)}, y_{(1)}), \ldots, (x_{(r+k-1)}, y_{(r+k-1)})\} & (2-k \leqslant r < 1), \\ \{(x_{(r)}, y_{(r)}), \ldots, (x_{(r+k-1)}, y_{(r+k-1)})\} & (1 \leqslant r \leqslant n-k+1), \\ \{(x_{(r)}, y_{(r)}), \ldots, (x_{(n)}, y_{(n)})\} & (n-k+1 < r \leqslant n), \end{cases}$$

so that each $\hat{\mu}_r(x)$ is a local fit obtained from a maximum of k sample points, with fewer points being used near the ends of the sample. Now define the sample fitted values to be

$$\hat{y}_{(i)} = k^{-1} \sum_{r=i-k+1}^{i} \hat{\mu}_r(x_{(i)}) \qquad (1 \leqslant i \leqslant n),$$

i.e. the fitted value at the sample point $(x_{(i)}, y_{(i)})$ is the average of the local fits $\hat{\mu}_r(x_{(i)})$, over the k values of r for which $(x_{(i)}, y_{(i)}) \in s(r)$. We shall refer to this as the 'k-fold local fit' of the base model $\hat{\mu}(x)$. Extreme cases are the onefold local fit which has $\hat{y}_{(i)} = y_{(i)}$ for all i, and thus involves no smoothing of the data, and the '∞-fold local fit', which reproduces the global fit of $\hat{\mu}(x)$ to the entire sample. As may easily be verified, if $\hat{\mu}(x)$ and its associated fitting method have the IBC property for $T = \Sigma_1^N a_i y_i$, the k-fold local fit retains that property, for any value of k.

In the simplest case where $\hat{\mu}(x)$ is constant, each $\hat{\mu}_r$ is simply the mean of the $y_{(i)}$ in $s(r)$. A simple, explicit matrix representation of the smoothing operation is then available: let M be the $(n + k - 1) \times n$ incidence matrix with elements

$$m_{ri} = \begin{cases} 1 & i \in s(r), \\ 0 & \text{otherwise}, \end{cases}$$

and let M^* be M with each row divided by its total, i.e.

$$m^*_{ri} = m_{ri}/m_{r\cdot};$$

then

$$\hat{\mathbf{y}}_s = k^{-1} M^{\mathrm{T}} M^* \mathbf{y}_s,$$

where $\mathbf{y}_s$ denotes the vector $(y_{(1)}, \ldots, y_{(n)})$, etc.

A k-fold local fit provides fitted values for the sampled units, but not for unsampled units as required for evaluation of $\hat{T}$. For this, interpolation is needed between the points $\{(x_{(1)}, \hat{y}_{(1)}), \ldots, (x_{(n)}, \hat{y}_{(n)})\}$, with extrapolation to any x-values outside the interval $[x_{(1)}, x_{(n)}]$. A simple prescription is to perform linear interpolation to calculate $\hat{\mu}(x_i)$ for x_i within any of the intervals $\{(x_{(r)}, x_{(r+1)}): r = 1, \ldots, n-1\}$, and to define $\hat{\mu}(x_i) = \hat{y}_{(1)}$ if $x_i < x_{(1)}$ and $\hat{\mu}(x_i) = \hat{y}_{(n)}$ if $x_i > x_{(n)}$. Any ties may be broken in an obvious way by averaging, for example. Other interpolation methods are of course possible, but our experience is that, at least with moderate or large sample sizes, the results from different interpolations are not discernibly different.

In Bennett (1994), a detailed theoretical and empirical investigation is made of estimators based on the k-fold local fit of various models, in particular the ratio model (8) for estimation of $T = N^{-1} \Sigma_1^N y_i$ under simple and stratified random sampling schemes.

Concerning the choice of k, our empirical experience is that, although a value such as $k \approx n/4$ is frequently better than $k = \infty$, there is often little or nothing to be gained by using much smaller values of k, except perhaps if the sample is sufficiently large that very localized features emerge strongly in the data. From the design-based viewpoint, the usual nonparametric smoothing concern about bias due to oversmoothing does not apply, since by construction $\hat{T}$ derived from a k-fold local fit is approximately design unbiased, regardless of k; and, as might be expected, the design variance of $\hat{T}$ is often found to be minimized by a fairly large choice of k. Of course, an asymptotic model-based analysis would still require $k \to 1$, i.e. the bandwidth tends to 0, as $n \to \infty$ for consistency under a completely nonparametric model (e.g. Dorfman (1992), Chambers *et al.* (1993) and Dorfman and Hall (1993)).

An alternative nonparametric regression approach would be via penalized likelihood fitting of a general function $\hat{\mu}(x)$ as described in, for example, Green and Silverman (1994), resulting in a natural cubic spline with knots at the sample values $x_{(r)}$. This method could presumably also be tuned to satisfy the IBC condition for a particular target quantity T. We have not

pursued this in detail but would be surprised if the results differed very much from those of a k-fold local fit offering a comparable degree of smoothing.

3. Examples

3.1. Estimating the size of a subdomain

Suppose that y_i indicates membership of a population subdomain, i.e. $y_i = 1$ if a member and $y_i = 0$ if not, and that a_i is the known size of each population unit i. Then $T = \Sigma_1^N a_i y_i$ is the total size of the subdomain. Application contexts include the two that were outlined briefly in Section 1, namely audit sampling and market research. In the following discussion, for concreteness, we use the terminology of a tax auditing application, but the methodology is quite general. For further details of the auditing context, see Stokes (1990).

In the tax auditing application introduced in Section 1, a_i is the size of the ith transaction and y_i is binary, with $y_i = 1$ if the transaction is taxable and $y_i = 0$ if not. If s is drawn by simple random sampling from the population of transactions, then as shown in Section 2.2.1 the maximum likelihood fit of a simple logistic regression

$$\hat{\mu}(a_i) = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 a_i)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 a_i)} \tag{13}$$

is IBC for $T = \Sigma_1^N a_i y_i$. Further terms can be added as necessary to the linear predictor to accommodate non-linearity, on the logit scale, of the dependence on a_i, or to allow dependence on other covariates if such are available.

Other possible IBC models for the regression of y_i on a_i here include the linear model

$$\hat{\mu}(a_i) = \hat{\beta}_1 + \hat{\beta}_2 a_i, \qquad \text{var}(y_i) = \text{constant}, \tag{14}$$

the ratio model

$$\hat{\mu}(a_i) = \hat{\beta}, \qquad \text{var}(y_i) \propto 1/a_i, \tag{15}$$

which when fitted by weighted least squares yields the standard ratio estimator

$$\hat{T}_{\text{ratio}} = \sum_1^N a_i \sum_s a_j y_j \Big/ \sum_s a_j,$$

and the 'expansion' model

$$\hat{\mu}(a_i) = \hat{\beta}/a_i, \qquad \text{var}(y_i) \propto 1/a_i^2, \tag{16}$$

which similarly yields the simple expansion estimator

$$\hat{T}_{\text{expand}} = (N/n) \sum_s a_i y_i.$$

Models (14)–(16) are included here largely for comparison. Model (16) may be re-expressed as $[E(a_i y_i) = \hat{\beta}, \text{var}(y_i) = \text{constant}]$, which is plainly unrealistic for the context. The ratio model (15), however, is more familiarly written as $[E(a_i y_i) = \hat{\beta} a_i, \text{var}(a_i y_i) \propto a_i]$ and is very

commonly used in the audit sampling context (Stokes, 1990), as it is in many others. One potential advantage of models (14)–(16) is that they each can be written in terms of a prediction formula that is linear in a_i, and thus they can still be used in situations where the individual values of a_i are unknown outside the sample but where the mean transaction size $\bar{a} = N^{-1} \Sigma_1^N a_i$ is known; the same cannot be said of a logistic regression, for example.

Fig. 1 displays the results of a small simulation experiment to explore the performance of various estimators under 2% simple random sampling from a synthetic population of 20000 transactions (a_i, y_i). The population, available electronically via the first author's home page at `www.stats.ox.ac.uk`, was designed to reproduce some features of real audit populations and also to be not very well described by standard models such as linear or logistic regressions; it was generated by sampling from the superpopulation model

$$a_i \sim \text{gamma}(4, 1), \qquad P(y_i = 1|a_i) = 0.9 \frac{\exp(a_i - 0.5)}{1 + \exp(a_i - 0.5)}. \tag{17}$$

'Expansion', 'ratio', 'linear' and 'logit–linear' in Fig. 1 are the estimators corresponding to the models (16), (15), (14) and (13) above. 'Logit–quadratic' results from adding a quadratic term $\hat{\beta}_2 a_i^2$ to the linear predictor in model (13), whereas 'logit–spline' is a two-parameter elaboration of model (13) which uses a piecewise linear fit on the logit scale, with knots at the 33rd and 67th percentiles of the distribution of a_i. Other design consistent estimators considered are two different k-fold fits of the expansion model (16), or equivalently of a simple mean model applied directly to the $\{a_i y_i\}$; the expansion estimator itself is thus the corresponding '∞-fold' fit. Also displayed in Fig. 1 are three estimators which depart systematically from the IBC condition. The first is based on a simple logistic regression as in model (13) but with $\log(a_i)$ replacing a_i as the covariate. The other two are based on simpler models: 'stratified' refers to the stratified mean model

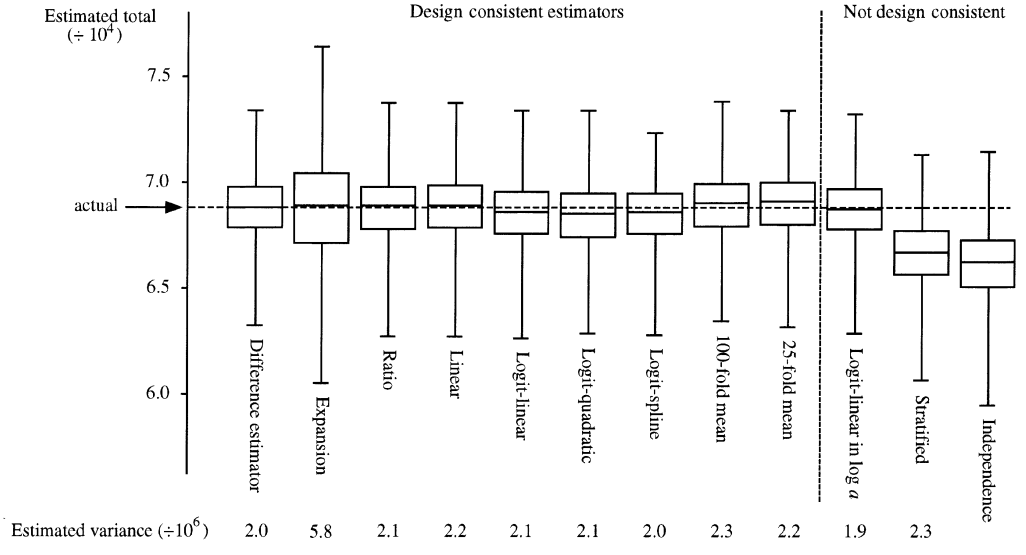

Fig. 1. Empirical sampling distribution of estimators of the size of a subdomain under simple random sampling

$$\hat{T}_{\text{strat}} = \sum_{i=1}^{N} a_i \hat{\mu}_g I_{ig},$$

where $g = 1, 2, 3$ are strata divided at the 33rd and 67th percentiles of a_i, I_{ig} indicates stratum membership and $\hat{\mu}_g$ is the sample mean of $y_i I_{ig}$; 'independence' refers to the unstratified version $\Sigma_1^N a_i (n^{-1} \Sigma_s y_i)$, naturally associated with a model in which the first two moments of y_i, at least, do not depend on a_i (see Stokes (1990), 'model η_0'). Finally in Fig. 1, included as a standard comparator, is the design consistent difference estimator (3) that would result from using as $\hat{\mu}(a_i)$ the 'true' superpopulation regression in model (17).

Fig. 1 is based on 2000 simple random samples of size 400. The design consistent estimators all have good bias properties, as expected. With the exception of the expansion estimator, which is based on the very unrealistic model (16), the variances of different methods are essentially the same. The standard error attached to each of the variance estimates shown is approximately 0.1×10^6. Comparison with the difference estimator of the true model, which would of course be unavailable in practice, indicates that as far as design variance is concerned there is little or nothing to be gained in this example by using more elaborate models than the models here which involve between two and four parameters.

Among the design inconsistent estimators, two miss the target by a considerable margin. The third, based on logistic regression of y_i on $\log(a_i)$, is almost unbiased in repeated sampling; this model does not have the IBC property, but $\Sigma_s a_i (\hat{y}_i - y_i)$, although not identically equal to 0, is found to have a mean in repeated sampling of only about 6, i.e. approximately 6/400 per unit sampled. Thus this estimator's bias, which would dominate asymptotically, is not serious in samples of moderate size.

To illustrate behaviour under an unequal probability sampling scheme, Fig. 2 displays the results of a second experiment involving stratified random sampling from the same population. Samples of size 190, roughly 1% of the population, were drawn by combining simple random subsamples of size 95 from each of two strata, the first stratum being the top 5% of transactions ranked by size, the second being the remaining 19000 smaller transactions. Stratum allocation here is thus highly non-proportional, with inclusion

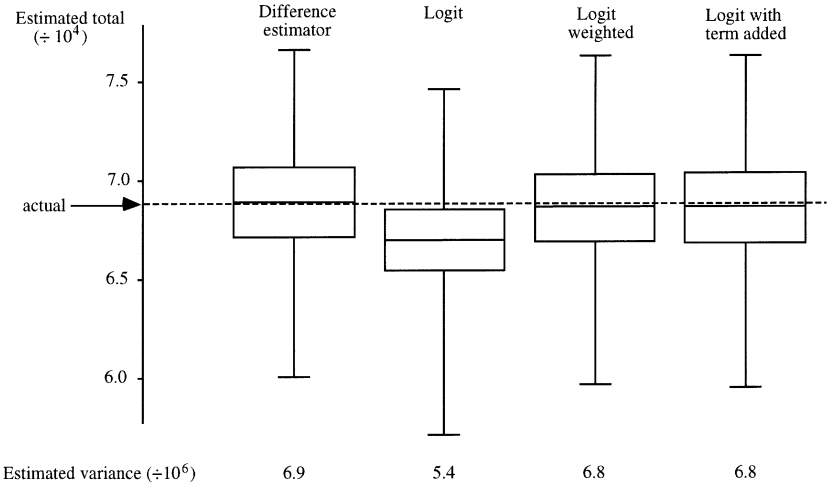

Fig. 2. Empirical sampling distribution of estimators of the size of a subdomain under a stratified sampling scheme

probabilities π_i 19 times greater for the largest 5% than in the remainder of the population. Fig. 2 is again based on 2000 samples and shows the empirical sampling distribution of four estimators:

(a) the difference estimator — the standard difference estimator, based on the true population-generating process $\hat{\mu}(a_i) = P(y_i = 1|a_i)$ as in model (17), and taking account of the unequal $\{\pi_i\}$;
(b) the logit estimator — $\hat{T}_{\text{pro}}$ (or, equivalently here, $\hat{T}_{\text{pre}}$) from the unweighted maximum likelihood fit of the simple logistic regression (13);
(c) the weighted logit estimator — $\hat{T}_{\text{pro}}$ from the fit of model (13) weighted by $w_{ir} = \pi_i^{-1}$, i.e. the pseudomaximum likelihood fit of model (13);
(d) the logit estimator with a term added — $\hat{T}_{\text{pro}}$ (or again, equivalently, $\hat{T}_{\text{pre}}$) from the unweighted maximum likelihood fit of the elaborated logistic regression

$$\hat{\mu}(a_i) = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 a_i + \hat{\beta}_3 a_i \pi_i^{-1})}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 a_i + \hat{\beta}_3 a_i \pi_i^{-1})},$$

which is piecewise linear on the logit scale since π_i is constant within strata.

Estimators (a), (c) and (d) are design consistent by construction, but estimator (b) is not. Fig. 2 shows that, whereas the logistic fit which takes no account of the sampling scheme yields an estimator with smaller variance than the other three methods, its bias dominates even at this relatively small sample size. Among the three design consistent estimators there is little difference in terms of performance in repeated sampling: the simple logistic regression (13), which is known here to be an incorrect description of the population, when equipped with the IBC property by either of the two devices used above is comparable with the 'ideal' difference estimator.

3.2. Estimating a bootstrap tail probability

In Monte Carlo sampling the regression modelling of simulation output to improve estimation efficiency is known as the method of control variates (e.g. Hammersley and Handscomb (1964) and Ripley (1987)). If the estimand is the population mean of y and the variable x has known population mean $\bar{x}$, say, a linear model $\hat{\mu}(x) = \hat{\beta}_1 + \hat{\beta}_2 x$ yields the natural projective estimator $\hat{\beta}_1 + \hat{\beta}_2 \bar{x}$. If not only the mean but also the entire distribution of x is known, and especially if that distribution is finite, the possibility exists of using a non-linear model, with potentially greater efficiency gains.

As an instance of this consider the estimation of a tail probability in the bootstrap distribution of some statistic; see, for example, Efron and Tibshirani (1993) for an introduction to bootstrap ideas. Here the population of interest is the finite collection of all possible resamples i from an observed sample distribution function, and $y_i = I(z_i > c)$, where z_i is the statistic of interest, c is a fixed constant and $I(\cdot)$ is the indicator function. Since y_i is binary, a natural style of design consistent model for estimating $\bar{y}$ under simple random sampling, either with or without replacement, is a logistic regression

$$\hat{\mu}(x_i) = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}, \tag{18}$$

in which $\hat{\beta}_1$ is present to ensure the IBC property and x_i is a suitable control variate. The essential requirement of x_i is that the mean of $\hat{\mu}(x_i)$, in the population of resamples, should be available analytically to enable its direct use as the projective estimator of the 'target'

$T = P(z_i > c)$. In the bootstrap context, candidate control variates typically include simple functions of the resample order statistics, or counts of the occurrence in resamples of some known probability event; in principle, x may be multivariate.

The most standard bootstrap sampling scheme is simple random sampling with replacement. Alternative, unequal probability sampling plans arise from the notion of importance sampling, in which the relative frequency of resample i is deliberately changed, say from f_i to g_i, to provide more information on the target quantity; see, for example, Hammersley and Handscomb (1964), Ripley (1987) or in the specific context of bootstrap methods Hesterberg (1996). Much of the development in Sections 1 and 2 continues to apply, and in particular the natural projective estimator as in equation (1) is consistent for the population mean of $a_i y_i$ if the IBC condition (4) holds, where now q_i denotes the ratio f_i/g_i of relative frequencies under simple random sampling and under the sampling plan actually used.

As a simple example, suppose that the empirical distribution function to be resampled is univariate with increments $1/m$ at each of m distinct points of support, and that the statistic of interest is the mean, i.e. z_i is the mean of resample i. Let x_i be the number of items in resample i which themselves exceed c. Then, in simple random resampling with replacement, the distribution of x_i is binomial(m, p), where p is the proportion of the support points which exceed c, and the mean of $\hat{\mu}(x_i)$ in equation (18) is straightforwardly calculated, whatever the values of $\hat{\beta}_1$ and $\hat{\beta}_2$. No claim of optimality is made for this particular choice of control variate, its primary motivation being its simplicity for illustration; it is reasonable, though, to expect this x_i to be sufficiently correlated with the resample mean z_i to be of some use in variance reduction.

Table 1 presents the results of a small experiment to compare the logistic regression (18) with the standard linear model approach to exploiting the control variate x_i, under simple random resampling. The 10 rows of the table correspond to 10 different distribution functions that were resampled, thus defining 10 different bootstrap populations of interest. Each of the 10 distribution functions has $m = 10$; they were in fact generated at random by simulation from the normal density, but for the present purposes their genesis is not very important. For each of the 10 distribution functions F_j, $j = 1, \ldots, 10$, a value of c_j was determined such that in resampling $T_j = P_{F_j}(z_i > c_j) \approx 0.025$, mirroring the design of a similar study by Efron and Tibshirani (1993), section 23.7. The performance of estimators of each T_j based on 1000 bootstrap resamples from F_j was then evaluated empirically using 100

Table 1. Empirical efficiency of estimators of an upper 2.5% tail probability, for 10 different bootstrap populations

Population	*Linear*	*Logit*
1	1.3	2.2
2	1.3	1.7
3	0.95	1.2
4	1.1	1.3
5	1.1	1.1
6	1.2	2.4
7	1.0	1.5
8	1.3	1.4
9	1.1	1.8
10	1.2	2.1
Average	1.2	1.7

replications of the whole procedure; Table 1 shows the estimated efficiency of each estimator relative to the simple resampling estimator $\Sigma_{i=1}^{1000} I(z_i > c_j)/1000$.

It might be expected *a priori* that the logistic regression (18) would be more realistic than a linear model for the binary response y_i here and therefore should yield a more efficient estimator. The results in Table 1 show that the logit model is consistently more efficient, offering an average improvement of about 70%, compared with the linear model's 20%. In both cases the bias for estimating $\bar{y}$ was found to be negligible; both of these canonical link models are IBC on account of the intercept term.

Generalization to importance sampling is straightforward, the IBC condition being satisfied if

$$\sum_s q_i \hat{\mu}(x_i) = \sum_s q_i y_i,$$

as may be achieved by using the $\{q_i\}$ as weights, for example. Hesterberg (1996) and Firth (1996) have demonstrated situations where the efficiency of importance sampling is substantially improved by the use of an appropriate control variate via a linear model; when y_i is binary, still greater improvements may be possible by using a logistic regression to exploit the control variate.

4. Concluding remarks

In this paper, the focus has been on models and estimators derived from them. For inference on T, estimation of $\text{var}(\hat{T})$ will also be needed. A detailed discussion is outside the scope of this paper, and here we note only that

(a) the variance to be estimated may be the variance in repeated sampling ('design variance') or variance under the assumed model ('model variance'), or a hybrid, depending on one's inferential standpoint, and
(b) for $\hat{T}$ derived from a parametric model, some general asymptotic methods are available, e.g. for model variance as in Valliant (1985) and for design variance as in Binder (1983).

Further work is needed on the estimation of $\text{var}(\hat{T})$ when $\hat{T}$ is derived from the sort of nonparametric model described in Section 2.4.

In Section 2 it was shown that a model without the IBC property may be equipped with it by the addition of an extra covariate, or by weighted fitting or by a combination of these two devices. The choice of method may in practice be constrained by features of the problem at hand; for example, weights need to be known only for the sample but in general to construct the estimator the value of an extra covariate must be known for the whole population. In terms of efficiency, there seems to be no clear cut 'best' recipe. Both devices are suboptimal under the model, and which one is more efficient will in general depend on the number of covariates already present and their relationship to any potential extra covariates. A simple, model-based principle might be that weights should always be chosen to maximize efficiency under the model; such weights would usually be determined by assumptions made about variances, leaving the addition of extra covariates as the favoured route to the IBC property. The addition of such extra, design-related covariates may also offer a degree of protection against model bias, as noted in Section 2 and by various other researchers (e.g. Little (1983), Rubin (1985) and Nordberg (1989)). An alternative principle, leading to a different conclusion, is optimality of the system of estimating equations used (Godambe and

Thompson, 1986), using a criterion defined with respect to both the model and repeated sampling. Optimum estimators based on this principle use design-based case weights throughout, to achieve the IBC property not only for T but also for the full set of parameters in the implicit model; this more restrictive framework thereby excludes the possibility of more than one route to the IBC condition for T alone and typically implies the use of weights which would be judged suboptimal by model-based criteria.

Finally, we note that estimators based on non-linear models will not be available in all sampling problems. In particular, if auxiliary variables $x_{i1}, \ldots, x_{ip}$ are available in the sample but only the totals $x_{.1}, \ldots, x_{.p}$ or means $\bar{x}_1, \ldots, \bar{x}_p$ are known for the population, $\hat{\mu}(\mathbf{x})$ is necessarily constrained to be linear to allow the estimator to be calculated.

Acknowledgements

The authors are grateful to C. J. Skinner, R. L. Chambers, R. B. Davies and the four referees for helpful comments on earlier versions of this work, and to S. L. Stokes for some motivating conversations. The second author was supported by a Science and Engineering Research Council research studentship at the University of Southampton.

References

Agresti, A. (1989) *Categorical Data Analysis*. New York: Wiley.

Bennett, K. E. (1994) A nonparametric regression approach to prediction in finite populations. *PhD Thesis*. University of Southampton, Southampton.

Binder, D. A. (1983) On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.*, **51**, 279–292.

Brewer, K. R. W. (1979) A class of robust sampling designs for large-scale social surveys. *J. Am. Statist. Ass.*, **74**, 911–915.

———(1995) Combining design-based and model-based inference. In *Business Survey Methods* (eds B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott), pp. 589–606. New York: Wiley.

Brewer, K. R. W., Hanif, M. and Tam, S. M. (1988) How nearly can model-based prediction and design-based estimation be reconciled? *J. Am. Statist. Ass.*, **83**, 128–132.

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., Dorfman, A. H. and Wehrly, T. E. (1993) Bias robust estimation in finite populations using nonparametric calibration. *J. Am. Statist. Ass.*, **88**, 268–277.

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

Dorfman, A. H. (1992) Non-parametric regression for estimating totals in finite populations. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 622–625.

Dorfman, A. H. and Hall, P. (1993) Estimators of the finite population distribution function using nonparametric regression. *Ann. Statist.*, **21**, 1452–1475.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.

Firth, D. (1996) Improved estimation after importance sampling. *Working Paper*.

Godambe, V. P. and Thompson, M. E. (1986) Parameters of superpopulation and survey population: their relationships and estimation. *Int. Statist. Rev.*, **54**, 127–138.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.

Hammersley, J. M. and Handscomb, D. C. (1964) *Monte Carlo Methods*. London: Chapman and Hall.

Hesterberg, T. C. (1996) Control variates and importance sampling for efficient bootstrap simulations. *Statist. Comput.*, **6**, 147–157.

Isaki, C. T. and Fuller, W. A. (1982) Survey design under a regression superpopulation model. *J. Am. Statist. Ass.*, **77**, 89–96.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, R. J. A. (1983) Estimating a finite population mean from unequal probability samples. *J. Am. Statist. Ass.*, **78**, 596–604.

Mantel, H. (1991) Making use of a regression model for inferences about a finite population mean. In *Estimating Functions* (ed. V. P. Godambe), pp. 217–221. Oxford: Clarendon.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Nordberg, L. (1989) Generalized linear modeling of sample survey data. *J. Off. Statist.*, **5**, 223–239.
Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.
Robinson, P. M. and Särndal, C. E. (1983) Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya* B, **45**, 240–248.
Royall, R. M. (1970) On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377–387.
Royall, R. M. and Herson, J. (1973) Robust estimation in finite populations: II, Stratification on a size variable. *J. Am. Statist. Ass.*, **68**, 890–893.
Rubin, D. B. (1985) The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 463–472. Valencia: Valencia University Press.
Särndal, C. E. (1980) On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, **67**, 639–650.
Särndal, C. E., Swensson, B. and Wretman, J. H. (1992) *Model Assisted Survey Sampling*. New York: Springer.
Särndal, C. E. and Wright, R. L. (1984) Cosmetic form of estimators in survey sampling. *Scand. J. Statist.*, **11**, 146–156.
Skinner, C. J., Holt, D. and Smith, T. M. F. (eds) (1989) *Analysis of Complex Surveys*. New York: Wiley.
Stokes, S. L. (1990) Estimating the size of a subdomain: an application in auditing. *J. Bus. Econ. Statist.*, **8**, 337–346.
Tibshirani, R. J. and Hastie, T. J. (1987) Local likelihood estimation. *J. Am. Statist. Ass.*, **82**, 559–568.
Valliant, R. (1985) Nonlinear prediction theory and the estimation of proportions in a finite population. *J. Am. Statist. Ass.*, **80**, 631–641.
Wolfe, R. (1996) Models and estimation for repeated ordinal responses, with application to telecommunications experiments. *PhD Thesis*. University of Southampton, Southampton.
Wright, R. L. (1983) Finite population sampling with multivariate auxiliary information. *J. Am. Statist. Ass.*, **78**, 879–884.