Scandinavian Journal of Statistics

Theory and Applications

Scandinavian Journal of Statistics, Vol. 37: 496–513, 2010
doi: 10.1111/j.1467-9469.2010.00691.x

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics. Published by Blackwell Publishing Ltd.

Transformation and Smoothing in Sample Survey Data

YANYUAN MA

Department of Statistics, Texas A&M University

ALAN H. WELSH

Centre for Mathematics and its Applications, Australian National University

ABSTRACT. We consider model-based prediction of a finite population total when a monotone transformation of the survey variable makes it appropriate to assume additive, homoscedastic errors. As the transformation to achieve this does not necessarily simultaneously produce an easily parameterized mean function, we assume only that the mean is a smooth function of the auxiliary variable and estimate it non-parametrically. The back transformation of predictions obtained on the transformed scale introduces bias which we remove using smearing. We obtain an asymptotic expansion for the prediction error which shows that prediction bias is asymptotically negligible and the prediction mean-squared error (MSE) using a non-parametric model remains in the same order as when a parametric model is adopted. The expansion also shows the effect of smearing on the prediction MSE and can be used to compute the asymptotic prediction MSE. We propose a model-based bootstrap estimate of the prediction MSE. The predictor produces competitive results in terms of bias and prediction MSE in a simulation study, and performs well on a population constructed from an Australian farm survey.

Key words: bootstrap, finite population prediction, model-based inference, non-parametric prediction, retransformation, transformation bias

1. Introduction

Consider a finite population of N units in which a survey variable Y has population values $Y_1, \ldots, Y_N$ and an auxiliary variable X has population values $X_1, \ldots, X_N$. For simplicity, we assume that X is a real variable; generalizations to vector X are reasonably straightforward. We assume that the values of X are known for all N population units, but that the values of Y are known only for a sample s of $n \leq N$ population units. Furthermore, the process used to select the population units to include in the sample is assumed to be conditionally independent of the values of Y given the values of X. Once the sample has been selected, the values of $Y_i, i \in s$ are known. The problem is to use $Y_i, i \in s$ and X_i, $i = 1, \ldots, N$ to predict the unknown finite population total $T = \sum_{i=1}^{N} Y_i$.

In this article, we propose a model-based transformation approach which incorporates smoothing and smearing to predict T by

$$\hat{T} = \sum_{i \in s} Y_i + n^{-1} \sum_{j \notin s} \sum_{i \in s} g[\hat{m}(X_j) + \{g^{-1}(Y_i) - \hat{m}(X_i)\}], \tag{1}$$

where g is a known transformation and $\hat{m}$ is either a parametric or a non-parametric estimator of the mean of the transformed survey variable $g^{-1}(Y)$. In the non-parametric case, $\hat{m}(x) = \sum_{k \in s} w_k(x) g^{-1}(Y_k)$, where $w_k(x)$ are weights (such as the Nadaraya–Watson, local linear or Gasser–Müller weights) constructed from a kernel function K with bandwidth $h > 0$. When g is the identity function, there is no need for smearing so the second term in (1) simplifies to $\sum_{j \notin s} \hat{m}(X_j)$. In the remainder of this section, we give an intuitive motivation for our use of (1) to predict T and then discuss its theoretical properties.

The model-based approach to predicting T (see, e.g. chapter 2 of Valliant *et al.*, 2000) is based on predictors of the general form

$$\hat{T}=\sum_{i\in s} Y_i+\sum_{j\notin s} \hat{Y}_j, \tag{2}$$

where $\hat{Y}_j$ is an estimator of $E(Y_j \mid Y_i, i\in s, X_1,\dots,X_N), j\notin s$. Under the usual assumption that the pairs (Y_i, X_i) are independent, $E(Y_j \mid Y_i, i\in s, X_1,\dots,X_N)=E(Y_j \mid X_j)$ and $\hat{Y}_j$ is taken to be the fitted value from a parametric regression model relating Y and X. The parametric regression model is commonly taken to be linear but it is convenient for later comparison to allow the more general formulation.

If the relationship between Y and X does not follow the assumed parametric form, it may be possible to transform Y so that the assumed parametric form holds on the transformed scale $g^{-1}(Y)$. That is,

$$g^{-1}(Y_i)=m(X_i)+\sigma\epsilon_i, \tag{3}$$

where $m(\cdot)=m(\cdot,\theta)$ is the regression function which is known up to the unknown regression parameter θ, σ is an unknown scale parameter and $\{\epsilon_i\}$ are i.i.d. with mean zero and variance one. (We also allow transformation of X but this does not introduce any bias so it is not necessary to make it explicit.) If $\hat{\theta}_g$ is an estimator of θ on the transformed scale, then we can predict $g^{-1}(Y_j)$ by $m(X_j,\hat{\theta}_g)$ and, after back transforming, predict Y_j by $\hat{Y}_j=g\{m(X_j,\hat{\theta}_g)\}$. This simple predictor is biased for Y_j because it estimates $g\{m(X_j,\theta)\}$ rather than $E[g\{m(X_j,\theta)+\sigma\epsilon\} \mid X_j]$. Even if the bias for predicting an individual Y_j is small when N and n are large, there are $N-n$ such terms in the predictor (2) so the bias accumulates and can make $\hat{T}$ severely biased. If we are prepared to assume an analytic form for the distribution of ϵ, we can try to adjust for this bias; for example, Karlberg (2000a,b) gives bias adjustments based on the log-transformation and the properties of the log-normal distribution. Alternatively, as pointed out by Chambers & Dorfman (2003b), we can use model calibration (Wu & Sitter, 2001) or smearing (Duan, 1983) to remove the transformation bias. We focus on smearing which estimates the expectation by an empirical average over the sample residuals, leading to the predictor $\hat{Y}_j=n^{-1}\sum_{i\in s} g[m(X_j,\hat{\theta}_g)+\{g^{-1}(Y_i)-m(X_i,\hat{\theta}_g)\}]$. This kind of prediction of individual observations is analysed in a very general (infinite population) context in Welsh & Zhou (2006). We have chosen to use smearing because it is conceptually simple, flexible enough to apply when m is non-parametric and fits into the model-based framework. On the contrary, implementing model calibration when m is non-parametric is awkward and as a model-assisted method which incorporates model information into design-based procedures, model calibration does not fit naturally into our model-based approach.

Chambers & Dorfman (2003b) point out that, in practice, both model calibration and smearing assume that the specified model fits the data on the transformed scale and can perform poorly when it does not. In particular, mis-specification of the regression function, a high number of zeros in the data and outliers can all lead to poor performance. The incorporation of a delta spike for handling zeros and robust estimation for handling outliers are treated in general in Welsh & Zhou (2006); in the finite population context, Chambers & Dorfman (2003b) suggested using robust estimates and incorporating robustness weights into the smearing predictor to deal with outliers. Both these ideas can be incorporated into the methodology we develop but, for simplicity, in this article, we consider only the less well-addressed issue of mis-specification. We adopt a non-parametric approach in which we treat m as a smooth function and estimate it non-parametrically. In (2) we use the smeared predictor $\hat{Y}_j=n^{-1}\sum_{i\in s} g\{\hat{m}(X_j)+g^{-1}(Y_i)-\hat{m}(X_i)\}$ to obtain (1). A general treatment of model-based non-parametric estimation of regression functions in finite population problems

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

is given by Chambers *et al.* (2003). Note that model-based non-parametric estimators are different and have different properties from design-based estimators of a smooth regression function such as considered by Breidt & Opsomer (2000).

In this article, we develop the prediction mean-squared error (MSE) properties of the parametric and non-parametric predictors without transformation and then with transformation and smearing. The calculations in each case proceed by writing the prediction error of the predictor $\hat{T}$ in (2) as

$$\hat{T}-T=\sum_{i\in s}Y_i+\sum_{j\notin s}\hat{Y}_j-\sum_{i=1}^{N}Y_i=P-\sum_{j\notin s}\{Y_j-E(Y_j\mid X_j)\},\tag{4}$$

where $P=\sum_{j\notin s}\{\hat{Y}_j-E(Y_j\mid X_j)\}$. Note that the two terms on the right-hand side are independent, because the first term involves only Y_i for i in sample, the second term involves only Y_i for i not in sample and the observations are independent. This independence between the two terms implies that the prediction MSE is

$$E(\hat{T}-T)^2=V+\sum_{j\notin s}\text{var}(Y_j\mid X_j),\tag{5}$$

where $V=E(P^2\mid X)=E([\sum_{j\notin s}\{\hat{Y}_j-E(Y_j\mid X_j)\}]^2\mid X)$. The second term on the right-hand side of (5) does not depend on the method of prediction so the problem is to evaluate V for different predictors. The calculations are fairly straightforward except in the case of the non-parametric predictor with transformation and smearing for which they are surprisingly complicated. Our main contribution is to obtain the leading terms in the expansion of $\sum_{j\notin s}\{\hat{Y}_j-E(Y_j\mid X_j)\}$ and to show that V in (5) is of order $N-n$ or, equivalently, of order n as $O(N)=O(n)$. (We will use $O(n)$ hereafter.) This is the same order as the second term in (5) so the prediction MSE is of order n in the parametric and non-parametric cases and even when smearing the predictions to remove transformation bias.

A practical issue with using non-parametric methods is that the (asymptotic) prediction MSE can be difficult to estimate. This is usually because of the bias terms – our result shows that the bias terms are of lower order than the leading stochastic terms so can be ignored when we estimate the asymptotic MSE. In this context however, some of the omitted terms may not be much smaller than the retained terms unless N and n are both very large. To overcome this difficulty, we propose in section 3 a model-based bootstrap procedure to estimate the prediction MSE.

The remainder of the article is organized as follows. The main results are presented in section 2. Estimation of the prediction MSE and inference are discussed in section 3 and the method is illustrated by a small simulation study in section 4. The article ends with a discussion in section 5, where the main features and findings of the proposed methods are recast and the potential limitations and possible extensions mentioned. The outline of the proof of the theory is deferred to an Appendix, and the technical details of the proof to the online Supporting Information available in connection with the online version of the article.

2. Prediction MSE results

Our main result gives the lowest order terms in the expansion of $\sum_{j\notin s}\{\hat{Y}_j-E(Y_j\mid X_j)\}$ for the non-parametric predictor with smearing. Simultaneously, we present results for the other predictors we have discussed because they are interesting in their own right, enable us to present a complete picture (including making comparisons) and provide context for the final result.

We assume throughout this section that $N \to \infty$ and $n \to \infty$ such that $0 < \lim n/N < 1$ and that $E\epsilon^4 < \infty$. Define $a(x, \epsilon) = g\{m(x) + \sigma\epsilon\}$, $b(x, \epsilon) = g'\{m(x) + \sigma\epsilon\}$ and $c(x, \epsilon) = g''\{m(x) + \sigma\epsilon\}$. Then, we assume the following functions exist and are continuous functions of x on the support of X:

$$\alpha(x) = Ea(x, \epsilon), \quad \beta(x) = Eb(x, \epsilon), \quad \gamma(x) = Ec(x, \epsilon),$$
$$Ea(x, \epsilon)^2,$$
$$Eb(x, \epsilon)^2, E\epsilon^2 b(x, \epsilon)^2,$$
$$Ec(x, \epsilon)^2, E\epsilon^4 c(x, \epsilon)^2.$$

These assumption allows us to bound various quantities on the support of X without requiring the transformation g and its derivatives to be bounded. This is useful because the common transformations are not bounded. Note that the Cauchy–Scharwz inequality enables us to bound other versions of these quantities, for example, as $E\epsilon^2 = 1$, $E\epsilon b(x, \epsilon) \le \{Eb(x, \epsilon)^2\}^{1/2}$, etc.

We present conditions for the two smeared predictors separately. Those listed in condition A are for the parametric model $m(x, \theta)$ and those in condition B are for the non-parametric model $m(x)$. The conditions for the predictors without transformation are a subset of these conditions.

Condition A

1. *The estimator $\hat{\theta}$ satisfies $\hat{\theta} - \theta = O_p(n^{-1/2})$.*
2. *The regression function $m(x, b)$ is differentiable in b, $m'(x, b)$ is uniformly continuous in b and $Em'(X, \theta) < \infty$.*
3. *The transformation g is monotone and differentiable and the derivative g' is uniformly continuous on any compact set.*

Condition (A1) allows for a wide class of estimators, including the widely used weighted least squares estimator and various robust estimators. Conditions (A2) and (A3) are smoothness conditions which allow us to expand both the regression function and the transformation when required.

For the non-parametric predictors, we consider linear estimators $\hat{m}(x) = \sum_{k \in s} w_k(x) g^{-1}(Y_k)$ of the regression function $m(x)$, where $w_k(x)$ are user-specified weights. Let K be a kernel function and $h > 0$ be the bandwidth. Then, common choices of weights include the Nadaraya–Watson weights

$$w_k(x) = \frac{K\{(x - X_k)/h\}}{\sum_{i \in s} K\{(x - X_i)/h\}},$$

the local-linear weights

$$w_k(x) = \frac{K\{(x - X_k)/h\}\{S_{n,2} - (X_i - x)S_{n,1}\}}{\sum_{i \in s} K\{(x - X_i)/h\}\{S_{n,2} - (X_i - x)S_{n,1}\}},$$

where $S_{n,m} = \sum_{i \in s} K\{(x - X_i)/h\}(X_i - x)^m$ (see Fan & Gijbels, 1996), or the Gasser–Müller weights

$$w_k(x) = \frac{1}{h} \int_{(X_{k-1} + X_k)/2}^{(X_k + X_{k+1})/2} K\{(x - u)/h\}\, du.$$

As we make explicit in the following conditions, any linear smoother whose weights satisfy standard conditions can be used in the predictor.

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

Condition B

1. *The linear estimator $\hat{m}(x)$ satisfies*

$$\hat{m}(x) - m(x) = h^2 C m''(x) + \sigma \sum_{k \in s} w_k(x)\epsilon_k + o(h^2) + O(n^{-1}).$$

2. *The weights $w_k(x)$ satisfy*

$$\sum_{i \in s} w_k(X_i) = O_p(1), \quad \sum_{j \notin s} w_i(X_j) = O_p(1),$$
$$\sum_{i \in s} w_i^2(X) = O_p\{(nh)^{-1}\}, \quad \sum_{i \in s} w_i^2(X_i) = O_p\{(nh^2)^{-1}\},$$
$$\sum_{j \notin s} \sum_{k \in s} w_k(X_j) w_k(X) = O_p(1), \quad \sum_{i \in s} \sum_{k \in s} w_k(X_i) w_k(X) = O_p(1),$$
$$\sum_{i \in s} w_i(X)\epsilon_i = O_p\{(nh)^{-1/2}\},$$

where the results hold uniformly with respect to the subindex and to X.

3. *The density p of X has compact support (so X is bounded) and $0 < c_1 < p(x) < c_2 < \infty$ on the support.*
4. *The regression function $m(x)$ is three times continuously differentiable.*
5. *The first two derivatives of the transformation g exist and the second derivative g'' is uniformly continuous on any compact set.*
6. *The bandwidth $h = O(n^{-\tau})$ with $1/4 < \tau < 1/2$.*

Conditions (B1) and (B2) are satisfied by the common choices of the estimator. Condition (B3) allows us to control the behaviour of the covariates and conditions (B4) and (B5) allow us to expand both the regression function and the transformation when required. In condition (B6), the lower bound $1/4 < \tau$ has to be strict to have the estimation variance dominate the bias, whereas the upper bound $\tau < 1/2$ can actually be relaxed to $\tau \leq 1/2$ without affecting the first order; this is discussed in more detail after the statement of the theorem. However, some terms are of order $h^{-1} = n^{\tau}$ which is the same order as the leading terms if $\tau = 1/2$. Thus, choosing $\tau < 1/2$ ensures that these terms are of smaller order than the leading terms and hence can be neglected asymptotically; in other words, choosing $\tau = 1/2$ does not change the order of the prediction MSE but adds extra terms to it. Note that (B6) excludes the usual $O(n^{-1/5})$ bandwidth because we are estimating the total which is an aggregate. The estimation variance is $O(n)$ which does not depend on h and allows us to select a relatively small bandwidth to eliminate the effect of the accumulated bias. Specifically, in calculating the total, the bias from the non-parametric regression estimator is accumulated to order nh^2. To contain the bias order within $n^{1/2}$, we have to exclude the usual $n^{-1/5}$ order from h. However, the relatively small bandwidth does not cause the variance order to deteriorate because of the aggregated nature of the total. This is reflected throughout the proof of the theorem.

Theorem

Suppose that either condition A or condition B holds as appropriate for the predictor. Then the first term $P = \sum_{j \notin s}\{\hat{Y}_j - E(Y_j \mid X_j)\}$ in the expansion (4) for each predictor is:

(i) *Parametric prediction without transformation*

$$P_1 = \sum_{j \notin s} m'(X_j, \theta)(\hat{\theta} - \theta).$$

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

(ii) *Non-parametric prediction without transformation*

$$P_2=\sigma\sum_{k\in s}\sum_{j\notin s}w_k(X_j)\epsilon_k.$$

(iii) *Parametric prediction with transformation and smearing*

$$P_3=\sum_{j\notin s}\beta(X_j)\{m'(X_j,\theta)-\frac{1}{n}\sum_{i\in s}m'(X_i,\theta)\}(\hat{\theta}-\theta)+\frac{1}{n}\sum_{j\notin s}\sum_{i\in s}\{a(X_j,\epsilon_i)-\alpha(X_j)\}.$$

(iv) *Non-parametric prediction with transformation and smearing*

$$P_4=\sigma\sum_{k\in s}\sum_{j\notin s}\beta(X_j)\{w_k(X_j)-\frac{1}{n}\sum_{i\in s}w_k(X_i)\}\epsilon_k+\frac{1}{n}\sum_{i\in s}\sum_{j\notin s}\{a(X_j,\epsilon_i)-\alpha(X_j)\}$$
$$-\frac{1}{n}\sigma\sum_{i\in s}\sum_{j\notin s}w_i(X_i)[b(X_j,\epsilon_i)\epsilon_i-E\{b(X_j,\epsilon_i)\epsilon_i\mid X_j\}].$$

The prediction MSE in each case is of order $O(n)$. That is, $E\{(\hat{T}-T)^2\}=O(n)$.

The proof is given in the Appendix.

The expression in (i) for the parametric predictor is familiar (see, e.g. Valliant *et al.*, 2000) and is included to establish a baseline. The expression in (ii) for the non-parametric predictor is similar to (i) because the bias contribution is of smaller order. As discussed in the 'Introduction', this occurs because of aggregation: we are predicting a sum rather than individual observations. Specifically, from (B1),

$$\sum_{j\notin s}\{\hat{Y}_j-E(Y_j\mid X_j)\}=\sum_{j\notin s}\{\hat{m}(X_j)-m(X_j)\}$$
$$=h^2C\sum_{j\notin s}m''(X_j)+\sigma\sum_{k\in s}\sum_{j\notin s}w_k(X_j)\epsilon_k+o_p(nh^2)+O_p(1)$$

so the leading term in the variance is

$$V=\sigma^2\sum_{k\in s}\left\{\sum_{j\notin s}w_k(X_j)\right\}^2=\sigma^2\sum_{k\in s}\sum_{j\notin s}\sum_{j'\notin s}w_k(X_j)w_k(X_{j'}),$$

which is of order n by condition (B2). When $h=O(n^{-\tau})$, the first term is of order $n^{1-2\tau}$, so the prediction bias is of smaller order than P_2 provided $\tau>1/4$ and hence the prediction MSE is of order n. It is also possible to choose $\tau=1/2$; this choice adds extra terms in (iv), although the final order of the prediction MSE remains $O(n)$. The advantage of choosing the boundary case $\tau=1/2$ is that it allows straightforward scaling when choosing bandwidth; this is described in detail in section 4. The transformation in (iii) affects the contribution of the parametric estimator through a multiplicative term $\beta(X_j)$, which comes from g', and by imposing centring on m'. The second term in (iii) captures the effect of smearing to remove transformation bias. The effect of this term is to increase the variability of the predictor. The transformation in (iv) has a similar effect to (iii) on the non-parametric estimator by introducing $\beta(X_j)$ and centring the weights. The effect of smearing here is to introduce two extra terms: one of these is the same as the smearing contribution in (iii) and the other combines the smearing with the estimator. As in (ii) there is no asymptotic bias and, as in (iii), the cost of smearing to remove bias is to increase variability. When the transformation g is the identity function, results in (iii) and (iv) reduce to those in (i) and (ii), respectively, in terms of their leading orders. The verification is straightforward and we omit the details here.

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

3. Inference

We can base inferences on the smeared non-parametric predictor by estimating the asymptotic prediction MSE and using a Gaussian approximation. However, this approach may not work very well because the asymptotic prediction MSE contains several terms and some of the omitted terms are only of slightly smaller order than those retained in the approximation. This suggests that it may be useful to investigate the use of the bootstrap to estimate the uncertainty in the predictor and to make inferences about the population total.

As the purpose of bootstrapping is to estimate the model-based prediction error or to set model-based confidence intervals for T, we construct a bootstrap distribution that is conditional on the sample s actually selected. We consider the approach suggested by Chambers & Dorfman (2003a), although other ways of bootstrapping may be possible. We treat the residuals

$$r_i = g^{-1}(Y_i) - \hat{m}(X_i) = \sigma\epsilon_i - h^2 C m''(X_i) - \sigma \sum_{k\in s} w_k(X_i)\epsilon_k + o(h^2) + O(n^{-1})$$

as estimates of the model errors. The residuals have conditional mean of order h^2 and approximate conditional variance $\sigma^2\{1+\sum_{k\in s} w_k(X_i)^2 - 2w_i(X_i)\} = \sigma^2 + O\{(nh)^{-1}\}$ so we can also use the standardized residuals

$$\frac{g^{-1}(Y_i) - \hat{m}(X_i)}{\sqrt{1+\sum_{k\in s} w_k(X_i)^2 - 2w_i(X_i)}}, \tag{6}$$

which have conditional mean of order h^2 and approximate conditional variance σ^2. The standardization produces a higher-order correction to the residuals and so can be omitted.

If we let $r_i^*, i=1,\ldots,N$ denote a simple random sample of residuals sampled with replacement from the sample residuals $r_i, i\in s$, then we can construct the bootstrap population

$$Y_i^* = g\{\hat{m}(X_i) + r_i^*\}, \quad i=1,\ldots,N.$$

Here we assume g is well defined at $\hat{m}(X_i)+r_i^*$. For commonly used transformations such as the log-transformation (so g is the exponential function) or the Box–Cox transformation, this is indeed the case. For this population, we compute the bootstrap population total T^* and use the same sample units s as before to construct the predictor $\hat{T}^*$ as defined in (2). The difference $\hat{T}^* - T^*$ is the prediction error based on the sample s for this bootstrap population and, repeating these calculations, we obtain the bootstrap distribution for the prediction error $\hat{T} - T$. The expected squared error of this bootstrap distribution is an estimate of the prediction MSE. Following Chambers & Dorfman (2003a), we can also use the quantiles of the bootstrap distribution (percentile method) to set confidence intervals for T.

The bootstrap described previously is attractively simple because it ignores the possible effects of bias on the residuals. There is no transformation bias because the residuals are constructed without any back transformation; this is related to the fact that the bootstrap population is constructed by predicting individual values Y_i rather than their conditional mean values $E(Y_i \mid X_i)$. [We do compute the bootstrap population total (an aggregated quantity) but it is more appropriate to treat this as the actual total of the bootstrap population (hence unsmeared) rather than as an estimate which can be smeared.] The residuals do include smoothing bias from estimating the regression function m; see Davison & Hinkley (1997, section 7.6). We can adjust for this bias if we can estimate C and m'' (and then consider modifying the approximate variance used in standardization to accommodate their uncertainty) but this introduces complications which we prefer to avoid. As we are estimating individual errors rather than aggregated errors, we should use a classical (larger) band-

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

width h to construct the residuals than we use to construct the non-parametric predictor of the population total.

4. Simulation

We illustrate the proposed method in a model-based simulation study. We performed 250 simulations in which samples of size $n=400$ were generated from populations of size $N=1600$ so the non-sample size is $N-n=1200$. We generated and kept fixed 400 X_is in the sample and 1200 X_is in the non-sample, but new Y_is were generated in each simulation. The X_is were generated from a uniform distribution $U(0,4)$, whereas the Y_is were generated from $g^{-1}\{m(X_i)+\epsilon_i\}$ with the ϵ_i from a standard normal distribution $N(0,1)$. We experimented with three different mean functions, $m(x)=50x/\{1+(x+4)^2\}, m(x)=2.5x/\{1+(x-1)^2\}$ and $m(x)=0.8(x-2)^2$, respectively, where the coefficients are selected so that the three functions have similar ranges. Finally, we used the identity transformation ($g^{-1}(Y)=Y$) and the log-transformation ($g^{-1}(Y)=\log(Y)$) as these seem to be the most widely used in practice.

We implemented the proposed method using the standard (Nadaraya–Watson) kernel estimator ($p1$) and the local linear estimator ($p2$) with the Epanechnikov kernel $K(x)=0.75(1-x^2)I(x^2<1)$ and with bandwidth $h_1=O(n^{-1/2})$ on the observed pairs $\{X_i, g^{-1}(Y_i)\}$, $i\in s$, to estimate the regression function $m(x)$. From condition (B6), $h_1=O(n^{-\tau})$ for any $1/4<\tau<1/2$ suffices. As we pointed out, using $h_1=O(n^{-1/2})$ instead of $o(n^{-1/2})$ changes the coefficient but not the order of the prediction MSE. In the bootstrap, we do not need to estimate the coefficients so the procedure carries through. The advantage of using the boundary order bandwidth is that it allows straightforward scaling between samples of different sizes in the cross-validation procedure described in the next paragraph. We denote the estimates of $m(x)$ by $\hat{m}_1(x)$ to emphasize the use of h_1. The predictors of the finite population total are then calculated through

$$\hat{T}=\sum_{i\in s} Y_i+\frac{1}{n}\sum_{i\in s}\sum_{j\notin s} g\{\hat{m}_1(X_j)+r_i\}.$$

To estimate the bias and variability of the proposed methods, we used a bootstrap method. We first obtained a new estimate of $m(x)$, denoted by $\hat{m}_2(x)$, from samples $\{X_i, g^{-1}(Y_i)\}$, $i\in s$, using bandwidth $h_2=O(n^{-1/5})$. Note that here h_2 has the same order as the usual optimal bandwidth for non-parametric regression as we use it to obtain individual residuals; see the discussion in the last paragraph of section 3. We then obtained the residuals $\tilde{r}_i=g^{-1}(Y_i)-\hat{m}_2(X_i), i\in s$, and formed bootstrap samples by constructing (X_k, Y_k^*), where $Y_k^*=g(X_k)+\tilde{r}_k^*, k=1,\dots,N$ and the $\tilde{r}_k^*$ were randomly selected with replacement from $\tilde{r}_i, i\in s$. The bootstrap procedure was repeated 500 times.

In practice we need to be able to select the two bandwidths h_1 and h_2. As h_2 optimizes the non-parametric fitting of $m(x)$ on the data $\{X_i, g^{-1}(Y_i)\}, i\in s$, it can be selected by the usual leave-one-out cross-validation. However, the selection of h_1 is not so straightforward. The final goal is to minimize the MSE of the predictor of the total when 1200 out of 1600 responses are missing so, ideally, we should use a leave-3/4-out cross-validation and then scale h_1 back by multiplying by $4^{1/2}$. Of course, it is not practical to calculate all $\binom{400}{100}$ cross-validation samples, so we select at random 50 sets of cross-validation samples and minimize the MSE of the predictor of the total over these 50 sets to obtain the optimal h_1. Although increasing 50 to a larger number will yield a closer approximation to the true optimal cross-validation bandwidth, the computational burden increases very quickly as well. Hence, we use this relatively small number of samples in the simulation.

We compared our method with several other methods including predictors based on fitting (a) a linear model ($aX+b$) with transformation and smearing; (b) a linear model through the origin (aX) without transformation; (c) a linear model without transformation; and (d) the expansion estimator of the total $\hat{T}=N/n\sum_{i\in s} Y_i$. To illustrate the importance of smearing when we transform, we also implemented 'no smearing' versions of the proposed estimators ($p1$), ($p2$) and method (a). The expansion estimator (d) is both a model-based estimator (for a simple homogeneous model) and a design-based estimator (for simple random sampling without replacement). We included two further design-based generalized regression (GREG) estimators of the total to complete the comparison between model- and design-based methods. We considered ($g1$) the GREG with constant variance and ($g2$) the GREG with variance proportional to X_i which is just the familiar ratio estimator. These estimators differ from their model-based counterparts (2) in how they treat the in-sample observations.

The simulation results are presented in Table 1 and Fig. 1. Several comments are worth making. (i) In all the situations, in terms of the sample MSE, ($p2$) has better performance than ($p1$). Hence, comparing the kernel and local linear estimators, the local linear estimator is usually preferred. Intuitively, this may be explained by the fact that the Nadaraya–Watson kernel estimator is a local constant estimator, which is generally outperformed by the local linear estimator in regions with sparse observations and at the boundary (see Fan & Gijbels, 1996, Härdle *et al.*, 2000, chapter 4). (ii) In most cases, the sample MSEs match the average of the estimated MSE rather well. This demonstrates that the inference derived in section 3 is valid and the asymptotic properties do not require huge sample sizes or bootstrap sizes. Here, sample MSE is the MSE of the estimators from the 250 simulation replicates compared with the true value, which is known in simulations, and average of the estimated MSE is the average of the 250 estimates of the MSE, each one calculated within a simulation replicated using only the sample information. Note that for the proposed estimators ($p1$) and ($p2$), the MSEs are dominated by the estimation variance. As we are estimating the total, the variance is N^2 times bigger than that of the mean. With $N=1600$ in the simulations, a variance of the order 10^6 is quite usual and does not indicate a problem. (iii) There are occasional situations, for example, for the model $\log(y)=2.5x/\{1+(x-1)^2\}$, in which the average of the estimated MSE does not match the sample MSE very closely. We conjecture that this may be caused by difficulties in bandwidth selection. In the bootstrap, to save computation time, we used the values of h_1 and h_2 chosen for the sample rather than re-estimating the bandwidths for each bootstrap sample. This may contribute to the observed performance because in the other experiments we conducted (not reported here) we found that if we use fixed bandwidths for h_1 and h_2, the difference between MSE and $\widehat{\text{MSE}}$ is rather small. Intuitively, choosing a bandwidth specific to each bootstrap sample contributes extra variability and hence will increase the total $\widehat{\text{MSE}}$, a phenomenon verified in a small experiment that we conducted. However, the estimated MSE is not always an underestimate so increasing the variability in the bootstrap is not always helpful. (iv) To examine the numerical usefulness of the asymptotic results derived in the theorem, we also calculated the estimated MSE for the estimator ($p1$) using the leading terms. As we expected, the results are too crude to be useful. In fact, for $m(x)=2.5x/\{1+(x-1)^2\}$, the estimated MSE is about 10 per cent of the sample MSE, whereas it falls to about 5 per cent for the other two mean functions. This confirms our intuition that the leading order is only slightly larger than the smaller terms that are ignored. It also suggests that the main contribution of the theorem is the evaluation of the rate of convergence.

Although our method is proposed in the model-based framework, out of curiosity, we also performed a simulation in the design-based framework. We still performed 250 simulations with the same sample size and non-sample size but the data generation procedure is different.

Table 1. *Simulation results. Replications* = 250, $n=400, N=1600$

	Bias	Vars	MSE	$\widehat{\text{MSE}}$	Bias	Vars	MSE	$\widehat{\text{MSE}}$
	$m(x)=50x/\{1+(x+4)^2\}, g^{-1}(y)=\log y$				$m(x)=50x/\{1+(x+4)^2\}, g^{-1}(y)=y$			
$(p1)$	2.79e2	4.17e6	4.24e6	5.09e6	−5.12	3.51e3	3.53e3	3.69e3
$(p2)$	3.88e2	3.84e6	3.99e6	4.01e6	−3.85	3.37e3	3.39e3	3.55e3
(a)	2.33e3	5.39e6	1.08e7	5.78e6	−1.92e1	3.38e3	3.74e3	4.03e3
(b)	−6.90e2	5.04e6	5.51e6	3.04e6	−3.03e2	2.77e3	9.43e4	3.85e3
(c)	−6.52e1	4.14e6	4.15e6	3.74e6	−1.92e1	3.38e3	3.74e3	4.05e3
(d)	−1.55e3	3.51e6	5.92e6	4.20e6	−7.05e1	3.36e3	8.34e3	6.01e3
$(g1)$	7.35e3	1.07e7	6.47e7	4.17e6	5.29e2	4.85e3	2.85e5	5.58e3
$(g2)$	1.67e2	4.90e6	4.93e6	4.55e6	2.32e1	3.59e3	4.12e3	6.41e3
$(p1n)$	−2.57e3	8.51e6	1.51e7	1.12e7	−5.87	5.25e3	5.28e3	5.31e3
$(p2n)$	−1.05e4	9.77e5	1.12e8	1.06e8	−4.39e1	4.07e3	6.00e3	5.03e3
(an)	−1.02e4	1.06e6	1.05e8	1.48e8	−1.92e1	3.38e3	3.74e3	4.03e3
	$m(x)=2.5x/\{1+(x-1)^2\}, g^{-1}(y)=\log y$				$m(x)=2.5x/\{1+(x-1)^2\}, g^{-1}(y)=y$			
$(p1)$	−3.95e2	1.84e6	2.00e6	1.59e6	−5.45	3.51e3	3.54e3	3.74e3
$(p2)$	−7.36e2	1.24e6	1.78e6	1.56e6	−6.27	3.38e3	3.42e3	3.55e3
(a)	2.70e2	1.83e6	1.90e6	2.15e6	−4.38e1	3.38e3	5.29e3	5.91e3
(b)	−4.68e3	7.52e5	2.26e7	1.81e6	−5.36e2	2.77e3	2.90e5	6.58e3
(c)	2.57e2	1.84e6	1.91e6	2.18e6	−4.38e1	3.38e3	5.29e3	5.93e3
(d)	3.25e2	1.88e6	1.99e6	2.19e6	−4.16e1	3.36e3	5.09e3	5.92e3
$(g1)$	−1.06e3	1.15e6	2.28e6	2.43e6	−5.42	4.85e3	4.88e3	9.54e3
$(g2)$	5.36e2	1.52e6	1.81e6	2.05e6	2.98e1	3.59e3	4.47e3	6.33e3
$(p1n)$	−1.74e3	4.32e6	7.34e6	4.67e6	−6.25	5.23e3	5.27e3	5.31e3
$(p2n)$	−6.47e3	3.53e5	4.22e7	3.78e7	−3.90e1	4.27e3	5.79e3	4.79e3
(an)	−8.61e3	1.25e5	7.43e7	7.63e7	−4.38e1	3.38e3	5.29e3	5.91e3
	$m(x)=0.8(x-2)^2, g^{-1}(y)=\log y$				$m(x)=0.8(x-2)^2, g^{-1}(y)=y$			
$(p1)$	−2.00e1	7.75e5	7.76e5	8.21e5	−3.20	4.13e3	4.14e3	3.97e3
$(p2)$	−2.23e2	4.88e5	5.38e5	5.42e5	−3.47	3.66e3	3.67e3	3.60e3
(a)	3.91e2	8.11e5	9.64e5	1.16e6	2.95e1	3.71e3	4.59e3	7.17e3
(b)	−1.74e3	8.77e5	3.92e6	9.91e5	−2.48e2	2.78e3	6.43e4	5.29e3
(c)	4.02e2	8.19e5	9.80e5	1.19e6	2.95e1	3.71e3	4.59e3	7.15e3
(d)	4.18e2	8.33e5	1.01e6	1.19e6	4.09e1	3.74e3	5.41e3	7.18e3
$(g1)$	4.20e2	1.50e6	1.68e6	1.26e6	9.96e1	5.06e3	1.50e4	8.03e3
$(g2)$	6.83e2	8.41e5	1.31e6	1.12e6	−1.43e1	3.42e3	3.63e3	6.58e3
$(p1n)$	−7.40e2	1.85e6	2.40e6	1.40e6	−6.47	5.49e3	5.53e3	5.90e3
$(p2n)$	−3.22e3	1.90e5	1.06e7	1.01e7	2.80e1	3.83e3	4.61e3	4.71e3
(an)	−5.33e3	2.63e5	2.85e7	3.30e7	2.95e1	3.71e3	4.59e3	7.17e3

Bias, sample bias; Vars, sample variance; MSE, sample mean-squared error; $\widehat{\text{MSE}}$, average of the estimated mean-squared error; $(p1)$, kernel estimator with transformation and smearing; $(p2)$, local linear estimator with transformation and smearing; (a), linear estimator with transformation; (b), linear regression through the origin without transformation; (c), linear regression without transformation; (d), expansion estimator; $(g1)$, GREG estimator with constant variance; $(g2)$, GREG/ratio estimator; $(p1n)$, $p1$ without smearing; $(p2n)$, $p2$ without smearing; (an), a without smearing.

Specifically, we generated a fixed population of 1600 (X_i, Y_i) pairs and then drew 250 independent simple random samples without replacement of 400 pairs from this population. All other aspects of the simulation design were identical to the model-based case. The simulation results are given in Table 2. We can see that although the proposed method is targeted at the model-based situation, its performance in the design-based situation is very good.

We also applied the proposed methods to samples drawn from the Australian Agricultural and Grazing Industries Survey (AAGIS) data from Kokic *et al.* (2000). This data set contains information on 1652 Australian broadacre farms. The variables we used are the total farm area in hectares (X) and the total cash costs of the farm during the survey year in Australian dollars (Y). Of the 1652 farms surveyed, 8 farms exhibited unusual data patterns and hence

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

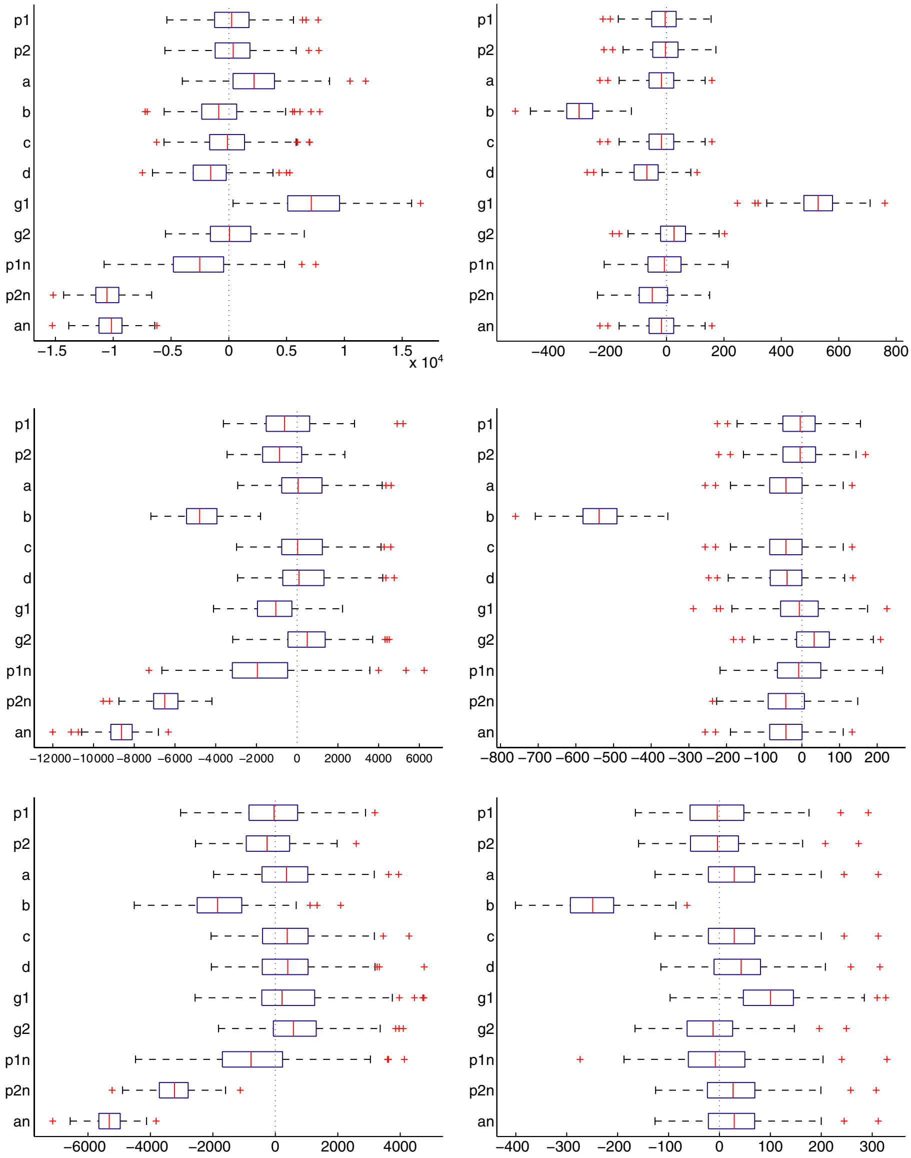

Fig. 1. Boxplots of the differences between the total estimates and the true totals for the cases presented in Table 1. The left column has $g^{-1}(y)=\log(y)$ and the right column has $g^{-1}(y)=y$. The top row has $m(x)=50x/\{1+(x+4)^2\}$, the middle row has $m(x)=2.5x/\{1+(x-1)^2\}$ and the bottom has $m(x)=0.8(x-2)^2$. The estimators are given in the same order as in Table 1.

were excluded from our analysis. The data are plotted in Fig. 2 on the raw scale and the log scale, respectively. Clearly, a log-transformation reveals the data pattern more clearly; so the log-transformation was used in our proposed method for the analysis.

We treated the 1644 farms as the population of interest and conducted a simulation study in which we selected 250 independent samples each of size 411 from the population. To simplify the computations, we fixed the bandwidth at $h=1$ for estimator ($p1$) and $h=3$ for estimator ($p2$). The choice of h is based on a preliminary cross-validation result, where we

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

Table 2. *Simulation results on design-based data. Replications* = 250, $n=400, N=1600$

	Bias	Vars	MSE	$\widehat{\text{MSE}}$	Bias	Vars	MSE	$\widehat{\text{MSE}}$
	$m(x)=50x/\{1+(x+4)^2\}, g^{-1}(y)=\log y$				$m(x)=50x/\{1+(x+4)^2\}, g^{-1}(y)=y$			
$(p1)$	−2.54e2	9.20e6	9.27e6	1.01e7	2.74	5.83e3	5.84e3	5.95e3
$(p2)$	−6.48e2	7.78e6	8.20e6	7.11e6	9.72e−1	5.82e3	5.82e3	5.75e3
(a)	1.20e3	8.23e6	9.67e6	8.40e6	1.87	6.24e3	6.25e3	6.26e3
(b)	−1.30e3	1.53e7	1.70e7	8.42e6	−3.29e2	5.89e3	1.14e5	6.23e3
(c)	−4.03e2	1.02e7	1.04e7	1.01e7	1.87	6.24e3	6.25e3	6.31e3
(d)	−4.46e2	1.07e7	1.09e7	1.07e7	1.74	8.17e3	8.18e3	8.77e3
$(g1)$	8.83e3	2.72e7	1.05e8	9.65e6	5.27e2	8.02e3	2.86e5	5.70e3
$(g2)$	2.37e2	1.09e7	1.10e7	9.19e6	2.14	8.64e3	8.65e3	6.09e3
$(p1n)$	−1.22e4	5.45e6	1.55e8	1.32e8	−1.13e1	9.28e3	9.41e3	9.01e3
$(p2n)$	−1.23e4	2.31e6	1.54e8	1.34e8	−3.95e1	6.08e3	7.64e3	7.28e3
(an)	−1.19e4	2.47e6	1.43e8	1.73e8	1.87	6.24e3	6.25e3	6.26e3
	$m(x)=2.5x/\{1+(x-1)^2\}, g^{-1}(y)=\log y$				$m(x)=2.5x/\{1+(x-1)^2\}, g^{-1}(y)=y$			
$(p1)$	−1.30e2	2.71e6	2.72e6	3.08e6	2.57	5.93e3	5.94e3	6.01e3
$(p2)$	−6.47e2	2.18e6	2.60e6	2.70e6	5.05e−1	5.94e3	5.94e3	5.74e3
(a)	−8.31e1	2.85e6	2.86e6	3.00e6	1.82	8.38e3	8.39e3	8.65e3
(b)	−5.38e3	1.85e6	3.08e7	2.84e6	−5.74e2	9.47e3	3.39e5	1.01e4
(c)	7.25e1	3.00e6	3.00e6	3.11e6	1.82	8.38e3	8.39e3	8.71e3
(d)	5.16e1	2.99e6	3.00e6	3.13e6	1.76e−1	8.43e3	8.43e3	8.74e3
$(g1)$	−1.68e3	2.71e6	5.52e6	2.65e6	−2.81e1	1.30e4	1.38e4	9.68e3
$(g2)$	−9.93e1	3.56e6	3.57e6	2.16e6	1.89	1.50e4	1.51e4	6.07e3
$(p1n)$	−7.44e3	1.38e6	5.67e7	5.90e7	−1.01	8.09e3	8.09e3	8.18e3
$(p2n)$	−7.78e3	9.11e5	6.15e7	5.79e7	−3.62e1	6.13e3	7.44e3	7.24e3
(an)	−9.43e3	6.56e5	8.96e7	8.80e7	1.82	8.38e3	8.39e3	8.65e3
	$m(x)=0.8(x-2)^2, g^{-1}(y)=\log y$				$m(x)=0.8(x-2)^2, g^{-1}(y)=y$			
$(p1)$	−1.80e2	1.59e6	1.62e6	1.85e6	2.47	6.18e3	6.18e3	6.14e3
$(p2)$	−1.17e1	1.17e6	1.17e6	1.39e6	1.54	5.90e3	5.90e3	5.86e3
(a)	1.46e2	2.15e6	2.17e6	2.23e6	−1.88	9.91e3	9.92e3	1.03e4
(b)	−2.69e3	9.62e5	8.17e6	1.78e6	−2.88e2	6.31e3	8.90e4	9.36e3
(c)	6.36e1	2.04e6	2.05e6	2.05e6	−1.88	9.91e3	9.92e3	1.04e4
(d)	9.88e1	2.02e6	2.03e6	2.06e6	1.40	9.87e3	9.87e3	1.04e4
$(g1)$	−3.26	1.28e6	1.39e6	1.56e6	5.01e1	1.01e4	1.26e4	8.77e3
$(g2)$	−4.45e1	1.91e6	1.91e6	1.34e6	2.03	1.08e4	1.08e4	7.20e3
$(p1n)$	−4.20e3	8.67e5	1.85e7	1.80e7	6.94	8.13e3	8.18e3	8.27e3
$(p2n)$	−3.73e3	5.20e5	1.44e7	1.58e7	7.10e1	6.25e3	1.13e4	1.04e4
(an)	−6.44e3	2.71e5	4.17e7	4.39e7	−1.88	9.91e3	9.92e3	1.03e4

Bias, sample bias; Vars, sample variance; MSE, sample mean-squared error; $\widehat{\text{MSE}}$, average of the estimated mean-squared error; $(p1)$, kernel estimator with transformation and smearing; $(p2)$, local linear estimator with transformation and smearing; (a), linear estimator with transformation; (b), linear regression through the origin without transformation; (c), linear regression without transformation; (d), expansion estimator; $(g1)$, GREG estimator with constant variance; $(g2)$, GREG/ratio estimator; $(p1n)$, $p1$ without smearing; $(p2n)$, $p2$ without smearing; (an), a without smearing.

averaged the selected bandwidths and rounded to the nearest integer. The results are presented in Table 3. Clearly, both proposed methods $(p1)$, $(p2)$ outperformed the other methods in terms of the MSEs, although the gain in comparison with method (a) is not substantial. A closer look at the right panel of Fig. 2 indicates that a non-parametric curve might not provide much better fit to the data than a linear function on the log scale, and this explains the similar performance of the three estimators. The estimated MSE is quite close to the sample MSE for both $(p1)$ and $(p2)$.

We performed similar studies on AAGIS data using the (cross-validation) bandwidth. In these studies, the performance gain of $(p1)$ and $(p2)$ and (a) is very similar, whereas all three continue to outperform the remaining estimators. The estimation of the MSE turns out to be less accurate for both $(p1)$ and $(p2)$.

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

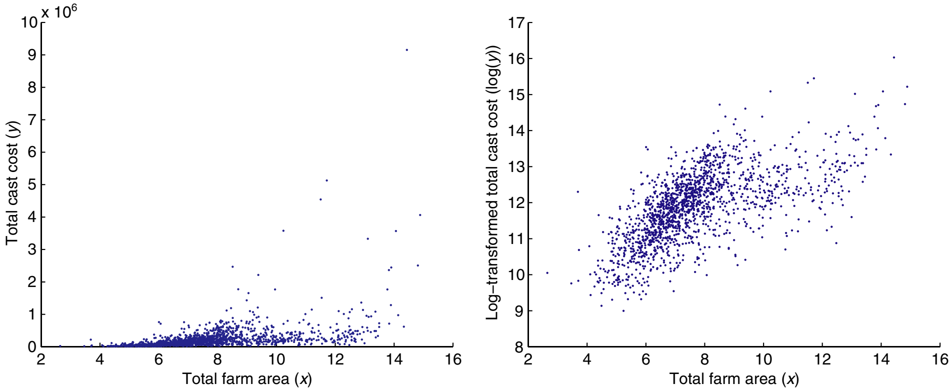

Fig. 2. Scatter plot of 1644 observations in the Australian Agricultural and Grazing Industries Survey data set, on the original scale (left panel) and the log scale (right panel), respectively.

Table 3. *Results for Australian Agricultural and Grazing Industries Survey data with eight outliers excluded. Results are based on* 250 *samples with sample size* $n=411$ *and non-sample size* $N-n=1233$

	Bias	Vars	MSE	$\widehat{\text{MSE}}$	Bias	Vars	MSE	$\widehat{\text{MSE}}$
(p1)	−5.49e3	4.78e8	5.08e8	4.66e8	−4.47e3	5.17e8	5.37e8	4.79e8
(p2)	−4.52e3	4.47e8	4.67e8	4.59e8	−5.18e3	5.22e8	5.49e8	7.85e8
(a)	6.32e3	4.41e8	4.81e8	4.19e8	–	–	–	–
(b)	3.99e4	1.05e9	2.64e9	7.65e8	–	–	–	–
(c)	1.76e3	5.56e8	5.59e8	7.25e8	–	–	–	–
(d)	3.86e3	6.14e8	6.29e8	9.03e8	–	–	–	–
(g1)	1.48e5	2.37e9	2.43e10	7.02e8	–	–	–	–
(g2)	1.70e3	7.49e8	7.52e8	6.25e8	–	–	–	–
(p1n)	−6.62e4	3.32e8	4.71e9	5.83e9	−6.82e4	4.46e8	5.10e9	3.63e8
(p2n)	−7.54e4	2.70e8	5.95e9	3.89e9	−7.63e4	3.83e8	6.20e9	4.36e9
(an)	−6.46e4	2.66e8	4.44e9	4.82e9	–	–	–	–

Bias, sample bias; Vars, sample variance; MSE, sample mean-squared error; $\widehat{\text{MSE}}$, average of the estimated mean-squared error; (p1), kernel estimator; (p2), local linear estimator; (a), linear estimator with transformation; (b), linear regression through the origin without transformation; (c), linear regression without transformation; (d), expansion estimator; (g1), GREG estimator with constant variance; (g2), GREG/ratio estimator; (p1n), p1 without smearing; (p2n), p2 without smearing; (an), a without smearing.

5. Discussion

In this article, we propose to estimate finite population totals using a transformation and smoothing approach with smearing to remove the bias caused by back transformation. The purpose of the transformation is to achieve additive homoscedastic error; in exchange, we may lose having an explicit parametric model for the mean. We have proved that even under a weak non-parametric model assumption, the prediction MSE can achieve the same order as under a parametric model. Moreover, the same result holds (so there is no further loss) under transformation with smearing. Hence, a non-parametric model does not cause any performance loss in terms of the order of the asymptotic error. The computation of the prediction MSE is nonetheless more challenging (as is always the case for non-parametric models), and we propose to use a bootstrap to obtain the estimated prediction MSE.

The result that non-parametric smoothing methods can achieve the same rate of convergence as parametric methods for predicting a finite population total is interesting and important. It is a consequence of the fact that we are predicting a sum or aggregate, which is like an integrated or expected regression function so parametric rates of convergence are achievable. In contrast, the familiar non-parametric rates are obtained when predicting a single observation. We can intuitively relate predicting aggregates versus predicting single observations to distribution function estimation versus density estimation. The distribution function is obtained by integration which, like summation, is a form of smoothing through aggregation. Non-parametric estimation of density functions is difficult and the best possible rates of convergence are the famous non-parametric ones. Non-parametric estimation of distribution functions is relatively easy and can be done at parametric rates. The reason is that the integration or aggregation has made the problem smoother and hence easier. In the survey context, predicting an individual observation is a difficult problem like the density estimation problem whereas predicting the total (an aggregate) is much easier, like the distribution function estimation problem. The possibility of achieving parametric rates of convergence from non-parametric predictors has been considered in the design-based case by Breidt & Opsomer (2000) but does not seem to have been discussed in the model-based sample survey literature. Hence, we give the result both with and without transformation for completeness.

Under transformation, smearing removes bias in the predictor but adds to its variability – an effect similar to that of the bias-adjustment approach discussed by Karlberg (2000a,b). However, as pointed out earlier, aggregating even small bias contributions can have a deleterious effect so it is important to remove the bias. Two sources of bias in our problem, namely smoothing and transformation, are reduced by aggregation and smearing, respectively. A third potential source of bias, model mis-specification, is reduced by our use of non-parametric smoothing. Nonetheless, although our model assumptions are much weaker than in parametric models, there are still assumptions and hence the potential for mis-specification effects. We assume throughout that the model is correct and do not consider bias caused by model mis-specification. Our results are practically important because they give a better understanding of the properties of non-parametric predictors in finite population problems and demonstrate a flexible method of prediction. In particular, our results show that non-parametric prediction is competitive with parametric prediction of a finite population total in terms of efficiency, convenience and is superior in terms of consistency and flexibility.

As the non-parametric approach should work for any smooth regression function, why not simply smooth the raw data and avoid transformation altogether? Hastie & Tibshirani (1990, section 7.1) provided a partial answer to this question. Although non-parametric smoothing weakens the assumptions on the model, it does not eliminate all of them and transformation may be employed to ensure that these assumptions are satisfied. When transformation induces homoscedastic errors, we can simplify the analysis because we do not need to also model a variance function. Transformation is also useful in practice for restricting predictions to an allowable range. Transformation can sometimes contribute to improving prediction performance, as we observe in our simulation study. Finally, when we have multiple covariates, an additive model is often used. In this case, transformation is often needed to achieve additivity. The analysis in the additive model inevitably builds on the simpler univariate case that we consider here. Our study of how the smoothing bias and transformation bias affect each other in the present simpler context, provides baseline results and a useful approach that can be extended in the future. In summary, the application of a nonlinear transformation g^{-1} to Y (often the logarithmic transformation) can be a useful way of achieving additive, homoscedastic errors. The purpose of making a transformation in a non-parametric model is different from that in parametric models (we change the error structure rather than the regression function) but the consequences in terms of prediction bias after back-transformation are the same and we need to extend the smearing technique to remove the bias.

It is worth pointing out that in practice, finding a suitable transformation g may not be easy. We can try to estimate the transformation within a parametric family of transformations g, such as the Box–Cox transformation, or a family of smooth transformations. However, estimating transformations and the non-parametric regression together gives rise to identifiability problems, so additional information, possibly in the form of instrumental variables will be needed; see, for example, Linton *et al.* (1997). For simplicity and focus, in this article, we assumed that a suitable transformation g is known.

In practice, the assumption that a transformation g can be found to achieve additive, homoscedastic errors should be checked using the sample observations. When such a g does not exist or is too hard to find, the method proposed here is not applicable. In such situations, we would recommend not performing any transformation. Instead, standard parametric or non-parametric regression should be performed on the sample data and the corresponding non-sample values should be predicted from these estimates. Straightforward summation of the non-samples should then be performed to estimate the total. Of course, inferences should then be modified to incorporate the actual error structure.

Finally, we have developed the method as a model-based approach but a design-based version can be constructed by incorporating sample inclusion probabilities appropriately. In our numerical experiment, we found that under simple random sampling without replacement (for which the sample inclusion probabilities are constant), the (model-based) procedure performed well as a design-based procedure. We caution that the theory and its derivation will be very different in the design-based framework, and a systematic study is needed before definite conclusions can be drawn.

Acknowledgement

This research was supported by the Australian Research Council (DP0559135) and a US NSF Grant DMS-0906341.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

More detailed proof of result (iv) in the theorem – non-parametric case with transformation and smearing.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Breidt, F. J. & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.* **28**, 1026–1053.

Chambers, R. L. & Dorfman, A. H. (2003a). Robust sample survey inference via bootstrapping and bias correction: the case of the ratio estimator. S3RI Methodology Working Papers, M03/13. Southampton Statistical Sciences Research Institute, University of Southampton.

Chambers, R. L. & Dorfman, A. H. (2003b). Transformed variables in survey sampling. S3RI Methodology Working Paper M03/21. Southampton Statistical Sciences Research Institute, University of Southampton.

Chambers, R. L., Dorfman, A. H. & Sverchkov, M. Yu. (2003). Non-parametric regression with complex survey data. In *Analysis of survey data* (eds R. L. Chambers & C. J. Skinner), 151–174. Wiley, New York.

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

Duan, N. (1983). Smearing estimate: a non-parametric retransformation estimate. *J. Amer. Statist. Assoc.* **78**, 605–610.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.

Härdle, W., Müller, M., Sperlich, S. & Werwaltz, A. (2000). *Non-parametric and semiparametric models*. Springer, New York.

Hastie, T. & Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall, London.

Karlberg, F. (2000a). Population total prediction under a lognormal superpopulation model. *Metron* **58**, 53–80.

Karlberg, F. (2000b). Survey estimation for highly skewed population in the presence of zeroes. *J. Off. Stat.* **16**, 229–241.

Kokic, P., Chambers, R. & Beare, S. (2000). Microsimulation of business performance. *Int. Stat. Rev.* **68**, 259–275.

Linton, O. B., Chen, R., Wang, N. & Härdle, W. (1997). An analysis of transformations for additive non-parametric regression. *J. Amer. Statist. Assoc.* **92**, 1512–1521.

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach*. Wiley, New York.

Welsh, A. H. & Zhou, X. H. (2006). Estimating the retransformed mean in a heteroscedastic two-part model. *J. Statist. Plann. Inference* **136**, 860–881.

Wu, C. & Sitter, R. R. (2001). A model calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* **96**, 185–193.

Received February 2009, in final form December 2009

Yanyuan Ma, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA.
E-mail: ma@stat.tamu.edu

Appendix: Proof of the theorem

We outline the proof of the theorem for case (iv) and then (iii). The full proof is available in the online Supporting Information which may be found in the online version of this article.

First note that the $\{\epsilon_i\}$ are identically distributed so that

$$\alpha(X_j)=n^{-1}\sum_{i\in s}E[g\{m(X_j)+\sigma\epsilon_i\}\mid X_j]=E[g\{m(X_j)+\sigma\epsilon_j\}\mid X_j]=E(Y_j\mid X_j).$$

Non-parametric case with transformation and smearing

Write

$$P=\sum_{j\notin s}\{\hat{Y}_j-E(Y_j\mid X_j)\}=\sum_{j\notin s}\{\hat{Y}_j-E(\hat{Y}_j\mid X)\}+\sum_{j\notin s}\{E(\hat{Y}_j\mid X)-E(Y_j\mid X_j)\}.$$

We develop an expansion for $\sum_{j\notin s}\hat{Y}_j$ which will also lead to an expansion for $\sum_{j\notin s}E(\hat{Y}_j\mid X)$. First, we make a (quadratic) Taylor series approximation to g in $\hat{Y}_j$ in powers of $\hat{m}(X_j)-\hat{m}(X_i)-m(X_j)+m(X_i)$ to obtain

$$\sum_{j\notin s}\hat{Y}_j=S_1+S_2+S_3+S_4,$$

where S_4 is the remainder term. Then, using the standard first-order bias plus stochastic term approximation to the estimator $\hat{m}$, we can write

$$\hat{m}(X_j)-\hat{m}(X_i)-m(X_j)+m(X_i)=h^2C\{m''(X_j)-m''(X_i)\}\\+\sigma\sum_{k\in s}\{w_k(X_j)-w_k(X_i)\}\epsilon_k+o(h^2)+O(n^{-1}).\qquad(7)$$

We substitute (7) into the remainder S_4, multiply out the quadratic terms and show that $S_4 = o_p(n^{1/2})$. Similarly, we also substitute (7) into S_2 and S_3 to obtain

$$S_2 = S_{21} + S_{22} \text{ and } S_3 = S_{31} + S_{32} + S_{33}.$$

Direct arguments lead to

$$\sum_{j \notin s} \hat{Y}_j = S_1 + S_{22} + S_{33} + o_p(n^{1/2})$$

and hence

$$\sum_{j \notin s} \{E(\hat{Y}_j \mid X) - E(Y_j \mid X_j)\} = E(S_{22} \mid X) + E(S_{33} \mid X) + o(n^{1/2}),$$

where

$$S_1 = n^{-1} \sum_{j \notin s} \sum_{i \in s} a(X_j, \epsilon_i),$$

$$S_{22} = n^{-1} \sigma \sum_{j \notin s} \sum_{i \in s} b(X_j, \epsilon_i) \sum_{k \in s} \{w_k(X_j) - w_k(X_i)\} \epsilon_k,$$

$$S_{33} = \frac{1}{2} n^{-1} \sigma^2 \sum_{j \notin s} \sum_{i \in s} c(X_j, \epsilon_i) \left[\sum_{k \in s} \{w_k(X_j) - w_k(X_i)\} \epsilon_k \right]^2.$$

In fact, the first term in the bias, $E(S_{22} \mid X) = O(nh^2) = o(n^{1/2})$, so we can include the bias in the remainder and write

$$P = \sum_{j \notin s} \{\hat{Y}_j - E(Y_j \mid X_j)\} = S_1 - ES_1 + S_{22} - ES_{22} + S_{33} - ES_{33} + o_p(n^{1/2}).$$

Now, rather straightforwardly, $S_1 - ES_1 = O_p(n^{1/2})$ but $S_{22} - ES_{22}$ and $S_{33} - ES_{33}$ involve three and four sums (so their variances involve six and eight sums), respectively, and are much more difficult to handle. We show that $S_{33} - ES_{33} = o_p(n^{1/2})$ but $S_{22} - ES_{22}$ can be written as a sum of two terms which are of order $O_p(n^{1/2})$ and other terms which are actually $o_p(n^{1/2})$. The leading terms given in (iv) in the theorem consist of $S_1 - ES_1$ and the two $O_p(n^{1/2})$ terms from $S_{22} - ES_{22}$ which are

$$n^{-1} \sigma \sum_{j \notin s} \left[\sum_{i \in s} \{b(X_j, \epsilon_i) - \beta(X_j)\} \right] \sum_{k \in s} \{w_k(X_j) \epsilon_k\}$$

and

$$-n^{-1} \sigma \sum_{j \notin s} \sum_{i \in s} w_i(X_i) [b(X_j, \epsilon_i) \epsilon_i - E\{b(X_j, \epsilon_i) \epsilon_i \mid X_j\}],$$

respectively.

Parametric case with transformation and smearing

The proof in the parametric case follows the same essential steps as in the non-parametric case but is simpler because there are no bias terms to control. Provided the transformation produces both the specified parametric form for the mean and additive, homoscedastic errors, direct Taylor expansion of g and m leads to

$$\sum_{j \notin s} \{\hat{Y}_j - E(Y_j \mid X_j)\} = S_1 + S_{21} + S_{22} + R,$$

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.

where

$$\begin{aligned}
S_1 &= n^{-1}\sum_{j\notin s}\sum_{i\in s}\{a(X_j,\epsilon_i)-\alpha(X_j)\},\\
S_{21} &= n^{-1}\sum_{j\notin s}\sum_{i\in s}\{b(X_j,\epsilon_i)-\beta(X_j)\}\{m'(X_j,\theta)-m'(X_i,\theta)\}(\hat\theta-\theta),\\
S_{22} &= \sum_{j\notin s}\beta(X_j)\{m'(X_j,\theta)-n^{-1}\sum_{i\in s}m'(X_i,\theta)\}(\hat\theta-\theta)
\end{aligned}$$

and, with $|\tilde\theta-\theta|\le|\hat\theta-\theta|$,

$$\begin{aligned}
R &= \Big|n^{-1}\sum_{j\notin s}\sum_{i\in s}[g'\{m(X_j,\theta)+\sigma\epsilon_i+m(X_j,\tilde\theta)-m(X_j,\theta)-m(X_i,\tilde\theta)+m(X_i,\theta)\}\\
&\quad\times\{m'(X_j,\tilde\theta)-m'(X_i,\tilde\theta)\}-b(X_j,\epsilon_i)\{m'(X_j,\theta)-m'(X_i,\theta)\}](\hat\theta-\theta)\Big|\\
&\le O_p(n^{1/2})\sup|g'\{m(X_j,\theta)+\sigma\epsilon_i+m(X_j,\tilde\theta)-m(X_j,\theta)-m(X_i,\tilde\theta)+m(X_i,\theta)\}\\
&\quad\times\{m'(X_j,\tilde\theta)-m'(X_i,\tilde\theta)\}-b(X_j,\epsilon_i)\{m'(X_j,\theta)-m'(X_i,\theta)\}|\\
&= o_p(n^{1/2}).
\end{aligned}$$

Also,

$$\begin{aligned}
&\operatorname{var}\left(n^{-1}\sum_{j\notin s}\sum_{i\in s}\{b(X_j,\epsilon_i)-\beta(X_j)\}m'(X_i,\theta)\mid X\right)\\
&= n^{-2}\sum_{i\in s}m'(X_i,\theta)^2\operatorname{var}\left(\sum_{j\notin s}\{b(X_j,\epsilon_i)-\beta(X_j)\}\mid X\right)\\
&= n^{-2}\sum_{i\in s}m'(X_i,\theta)^2\sum_{j\notin s}\sum_{j'\notin s}[E\{b(X_j,\epsilon_i)b(X_{j'},\epsilon_i)\mid X\}-\beta(X_j)\beta(X_{j'})]\\
&= O(n).
\end{aligned}$$

So $S_{21}=O_p(1)$ and the terms given in (iii) in the theorem consist of S_1 and S_{22}, which are both $O_p(n^{1/2})$.

© 2010 Board of the Foundation of the Scandinavian Journal of Statistics.