Statistical Science
2017, Vol. 32, No. 2, 190–205
DOI: 10.1214/16-STS589

Model-Assisted Survey Estimation with Modern Prediction Techniques

F. Jay Breidt and Jean D. Opsomer

Abstract. This paper reviews the design-based, model-assisted approach to using data from a complex survey together with auxiliary information to estimate finite population parameters. A general recipe for deriving model-assisted estimators is presented and design-based asymptotic analysis for such estimators is reviewed. The recipe allows for a very broad class of prediction methods, with examples from the literature including linear models, linear mixed models, nonparametric regression and machine learning techniques.

Key words and phrases: Machine learning, nonparametric regression, nearest neighbors, neural network, regression trees, survey asymptotics.

1. INTRODUCTION TO DESIGN-BASED ESTIMATION AND INFERENCE

The basic problem in survey statistics is estimation of characteristics of a target finite population. These characteristics can take many forms, but quantitative summaries such as means, totals, distribution functions and quantiles of variables of interest are most common. There are almost always many variables of interest. For instance, what is the average number of red snapper caught during an off-shore fishing trip in the Gulf of Mexico during 2014? What is the average weight of the red snapper caught? The average number of discarded red snapper? Same questions for dozens of other species of fish that might be caught. As another example, what is the total number of women with undergraduate degrees in statistics in the US in 2015? The percentage of such women who are in the labor force? The income of such women who are employed? What about women with graduate degrees in statistics? What about men, and what about dozens of other academic fields?

Real finite populations like the target populations in these examples are highly complex and heterogeneous, as are the response variables to be studied for those populations. There is therefore understandable reluctance to specify statistical models for the behavior of all the variables of interest in the population. Instead of relying on statistical modeling variable-by-variable for estimation and inference, design-based survey statistics uses *randomization* as the tool to select which population units to measure, and then constructs estimators that rely on this randomization for their statistical validity.

An immediate implication of relying on randomization instead of on an underlying stochastic population structure is that the values of the variables of interest in the population are treated as fixed but unknown quantities. Let y_k denote the nonrandom value of a variable of interest for the kth element in the finite population $U = \{1, 2, \ldots, N\}$. We focus on estimation of the finite population total $t_y = \sum_{k \in U} y_k$. More complex finite population parameters can often be written as explicit functions of finite population totals, like means, proportions, ratios and regression coefficients, over the whole population or over domains (subpopulations). Such explicit nonlinear functions can be estimated by plugging in appropriate estimators of the component totals. Other finite population parameters may be implicitly defined, as the solution of population-level estimating equations. In this case, the estimating equation can often be expressed as an explicit function of finite population totals; estimation of the estimating equation then leads to an estimation method for the implicitly-

F. Jay Breidt is Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877, USA (e-mail: jbreidt@stat.colostate.edu). Jean D. Opsomer is Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877, USA (e-mail: jopsomer@stat.colostate.edu).

defined finite population parameters. Hence, it is appropriate to focus on the estimation of the finite population total t_y for now, because it can be viewed as a fundamental building block of many other survey estimators of interest.

A second implication of the reliance on randomization in the design-based approach is that the randomness in the estimators is due only to the random selection of the sample. The *sampling design*, denoted $p(s)$, is a probability distribution on the set of all 2^N possible subsets of U; that is, $p(s)$ is the probability of selecting the particular sample, s. Define the *sample membership indicator* $I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise. For $k, \ell \in U$, let $\pi_k = \mathrm{E}[I_k] = \mathrm{P}[k \in s] = \sum_{s \subset U: k \in s} p(s)$ denote the *first-order inclusion probabilities* of the design and let $\pi_{k\ell} = \mathrm{E}[I_k I_\ell] = \mathrm{P}[k, \ell \in s] = \sum_{s \subset U: k,\ell \in s} p(s)$ the *second-order inclusion probabilities*. The design $p(s)$ is a *probability sampling design* if $\pi_k > 0$ for all $k \in U$.

The sampling design controls the random behavior of the sample, and hence of any estimators computed from it. Especially when the sampling design is complex, that is, including one or several levels of random selection with possibly unequal probabilities, it makes sense that incorporating design information in the construction of estimators is important, for both statistical validity and efficiency reasons. For any probability sampling design, the Horvitz and Thompson (1952) estimator incorporates design information via inverse-probability weighting,

$$\mathrm{HT}(y) = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} y_k \frac{I_k}{\pi_k}, \tag{1}$$

and is design-unbiased for t_y in the sense that, averaging over all possible samples,

$$\mathrm{E}\big[\mathrm{HT}(y)\big] = \sum_{k \in U} y_k \frac{\mathrm{E}[I_k]}{\pi_k} = t_y.$$

The variance of the Horvitz–Thompson estimator then depends on the covariance structure of $\{I_k\}_{k \in U}$,

$$\begin{aligned} \mathrm{Var}\big(\mathrm{HT}(y)\big) &= \sum_{k,\ell \in U} \mathrm{Cov}(I_k, I_\ell) \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \\ &= \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \end{aligned} \tag{2}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$. If $\pi_{k\ell} > 0$ for all $k, \ell \in U$, the design is said to be *measurable*, and the design variance (2) admits an unbiased estimator,

$$\widehat{\mathrm{V}}\big(\mathrm{HT}(y)\big) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}. \tag{3}$$

Because many survey estimators are smooth functions of weighted sums like (1) or related estimators, confidence intervals and hypothesis tests are typically conducted by appealing to asymptotic normality. In Section 3 and Section 5 below, we further describe a theoretical framework under which asymptotic properties of survey estimators are most often obtained.

The availability of an unbiased estimator of the finite population total and an unbiased estimator of its variance under any measurable probability sampling design, for any response variable of interest, and without the need to specify a model for the data, makes the design-based approach a simple and robust all-purpose statistical framework for survey statisticians. Once a sampling design is decided upon, an associated inverse-probability-weighting estimation procedure is available that will lead to estimators with quantifiable statistical properties. However, a disadvantage of the design-based approach is that the resulting estimators can be inefficient, sometimes dramatically so. This efficiency depends on the relationship between the population characteristics and the sampling design.

Improving the efficiency of estimators has been a major research focus within survey statistics, as well as the source of some controversy within the discipline. One possible approach to improving the efficiency of survey estimators is to use sampling designs that are carefully crafted to lead to efficient design-based estimators. In many situations where information about the target population is available prior to sampling, this can indeed lead to efficiency improvements. Another approach is to abandon the design-based paradigm altogether and instead postulate and fit a statistical model for the population variables of interest. To the extent that the specified model is correct, this approach can lead to estimators with better statistical properties than the design-based estimators, at the expense of additional effort in model validation and checking for each variable under study. A final approach, and the main focus of this article, is to incorporate additional population information and modeling into the design-based approach, to improve the efficiency of estimators while also maintaining the desirable design-based properties of approximate unbiasedness and consistency. This approach is often referred to as *model-assisted*, because it uses models to improve the efficiency of estimation, while remaining within the design-based inferential framework.

A unifying framework for the study of many model-assisted estimators is the *calibration* approach (Deville

and Särndal, 1992), in which sample weights are constructed to reproduce known population-level information, while remaining as close as possible (under some metric) to the original inverse-probability weights. See Särndal (2010) for a recent, comprehensive review of calibration estimation. We do not attempt to review this extensive literature, but focus instead on model-assisted estimators that are directly motivated by prediction ideas. Many such estimators, described below, have at least some calibration properties, but calibration to external controls is not the primary motivation in their construction.

2. AUXILIARY INFORMATION AND THE DIFFERENCE ESTIMATOR

In many survey estimation settings, *auxiliary information* is available at the population level. This information can take many forms and come from different sources. For instance, a recreational fishing license registry will have data on each registered angler in a state and might also have information on whether each angler owns a boat. A satellite image will have each pixel in a landscape classified as "forested," "not forested" or "undetermined." The information might be available for every unit in the population, or only in summary form, such as totals or means for the population.

We will write $\boldsymbol{x}_k$ for a vector of auxiliary variables, and at a minimum, we will assume that population totals $t_{\boldsymbol{x}} = \sum_{k\in U} \boldsymbol{x}_k$ are known and sample vectors $\{\boldsymbol{x}_k\}_{k\in s}$ are observed. Some estimators require the stronger condition that the individual vectors $\{\boldsymbol{x}_k\}_{k\in U}$ are known for the entire finite population, a setting that we assume for the moment. Suppose further that we have a "method" $m(\cdot)$ for predicting y_k from $\boldsymbol{x}_k$:

$$y_k \simeq m(\boldsymbol{x}_k),$$

subject to the condition that the method $m(\cdot)$ does not depend on the sample. For example, some studies alternate between regular surveys and periodic censuses, and the condition would be met for a method that relies only on updating a census value without reference to sample data.

Given the output of the method, the *difference estimator* of t_y is

$$\begin{aligned}\mathrm{DIFF}(y;m) &= \sum_{k\in U} m(\boldsymbol{x}_k) + \sum_{k\in s} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \\ &= \sum_{k\in U} m(\boldsymbol{x}_k) + \mathrm{HT}(y-m).\end{aligned} \tag{4}$$

This estimator is exactly unbiased, regardless of the quality of the method, since

$$\begin{aligned}\mathrm{E}\big[\mathrm{DIFF}(y;m)\big] &= \sum_{k\in U} m(\boldsymbol{x}_k) + \mathrm{E}\big[\mathrm{HT}(y-m)\big] \\ &= t_m + t_y - t_m = t_y.\end{aligned}$$

Because $\sum_{k\in U} m(\boldsymbol{x}_k)$ does not depend on the sample, it is not random, and the design variance of the difference estimator follows immediately from that of the HT estimator:

$$\begin{aligned}&\mathrm{Var}\big(\mathrm{DIFF}(y;m)\big) \\ &\quad= \mathrm{Var}\big(\mathrm{HT}(y-m)\big) \\ &\quad= \sum_{k,\ell\in U} \Delta_{k\ell} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell}.\end{aligned} \tag{5}$$

We therefore expect (5) to be smaller than the variance of $\mathrm{HT}(y)$ in (2) provided that the "residuals" $\{y_k - m(\boldsymbol{x}_k)\}$ have smaller variation than the "raw values" $\{y_k\}$. Whether or not the predictive method is good, the difference estimator will behave like the HT estimator. As above, under a measurable sampling design, the unbiased variance estimator for the difference estimator is

$$\begin{aligned}&\widehat{V}\big(\mathrm{DIFF}(y,m)\big) \\ &\quad= \sum_{k,\ell\in U} \Delta_{k\ell} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}},\end{aligned}$$

again without regard to the quality of the method. Under mild conditions (to be described in Section 3 below), the difference estimator will inherit consistency and asymptotic normality from the corresponding results for the HT estimator, provided only that the residuals $\{y_k - m(\boldsymbol{x}_k)\}$ are sufficiently well behaved.

While the difference estimator has obvious appeal as an exactly unbiased estimator with potentially reduced variance, it is rarely applied directly in practice, because it is quite rare to have access to a sample-independent method $m(\cdot)$ that gives good predictions of y_k. It is more natural to estimate $m(\cdot)$ based on sample data $\{(\boldsymbol{x}_k, y_k)\}_{k\in s}$.

3. SURVEY ASYMPTOTICS I

Before we discuss replacing $m(\cdot)$ by specific sample-based estimators, we describe a theoretical framework for asymptotic analysis of design-based estimators. Because we are selecting a random sample from a finite population, asymptotic arguments start from a sequence of finite populations U_N of size N, with

$N \to \infty$. With this sequence is associated a sequence of sampling designs $p_N(\cdot)$. For each N in the sequence, a sample $s_N \subset U_N$ is drawn according to design $p_N(\cdot)$, with sample size n_N. Inclusion probabilities $\pi_{kN}, \pi_{k\ell N}$ are associated with the design $p_N(\cdot)$. Following customary notational practice, the subscript N will be suppressed in all these quantities whenever possible.

Because the population and the design change with N, regularity conditions are needed on both to ensure that asymptotic results are well-defined. Considering the simplest estimator, $\mathrm{HT}(y)$, the following are examples of such regularity conditions for a design of nonrandom size n:

- D1. As $N \to \infty$, $nN^{-1} \to \pi^* \in (0, 1)$. For all N, $\min_{k \in U} \pi_k \geq \lambda_1 > 0$ and
$$\limsup_{N \to \infty} n \max_{k,\ell \in U : k \neq \ell} |\Delta_{k\ell}| < \infty.$$
- D2. The study variables $\{y_k\}_{k \in U}$ satisfy
$$\limsup_{N \to \infty} N^{-1} \sum_{k \in U} y_k^2 < \infty.$$

Condition D1 on the sequence of sampling designs $p_N(\cdot)$ is expressed in terms of its inclusion probabilities, the most common approach unless a specific sampling design is being considered (e.g., stratified simple random sampling). The lower bound on the π_k implies that the design is a probability sampling design, and the condition on the $\Delta_{k\ell}$ states that the dependence between sample membership indicators is sufficiently small. These conditions are satisfied for many classical sampling designs, including simple random sampling with and without replacement and their stratified versions, and also allow for unequal probability sampling and random sample size, in which case n above denotes the expected sample size, $\mathrm{E}[\sum_U I_k]$. Condition D2 on the sequence of finite populations U_N ensures that second-order finite population moments of the variables of interest have well-defined limits, a very mild condition.

These conditions are sufficient to show directly that

$$\begin{aligned} &\mathrm{Var}\big(N^{-1}\,\mathrm{HT}(y)\big) \\ &\quad \leq \frac{1}{N\lambda_1} \sum_{k \in U} \frac{y_k^2}{N} + \frac{\max_{k,\ell \in U : k \neq \ell} |\Delta_{k\ell}|}{\lambda_1^2} \\ &\qquad \cdot \left(\sum_{k \in U} \frac{|y_k|}{N} \right)^2, \end{aligned} \tag{6}$$

and that this bound converges to zero as $N \to \infty$. Taken together with the unbiasedness of $\mathrm{HT}(y)$, this implies design mean square consistency of HT and hence consistency with respect to the sequence of sampling designs $p_N(\cdot)$, that is, design consistency.

Further conditions are needed for inference. More specifically, sufficient conditions on the sequence of populations and associated designs are required to obtain the asymptotic normality of $\mathrm{Var}(\mathrm{HT}(y))^{-1/2}\{\mathrm{HT}(y) - t_y\}$ and consistent estimation of $\mathrm{Var}(\mathrm{HT}(y))$. The asymptotic normality of the Horvitz–Thompson estimator is often assumed explicitly, because sufficient conditions that hold for arbitrary designs are actually difficult to state. For specific designs, asymptotic normality results are available in the literature, including the classical result by Hájek (1960) for Poisson sampling and simple random sampling without replacement. Additional central limit theorems for stratified sampling include Krewski and Rao (1981), who considered stratified unequal probability samples with replacement, Bickel and Freedman (1984), who considered stratified simple random sampling without replacement, and Breidt, Opsomer and Sanchez-Borrego (2016), who considered general unequal probability designs, with or without replacement.

For consistent estimation of $\mathrm{Var}(\mathrm{HT}(y))$, we can proceed as we did above for the consistency of $\mathrm{HT}(y)$ itself. In addition to D1, we require

- D3. For all N, $\min_{k,\ell \in U_N} \pi_{k\ell} \geq \lambda_2 > 0$.

to ensure that the design is measurable, which in turn guarantees that $\widehat{\mathrm{V}}(\mathrm{HT}(y))$ in (3) is unbiased for $\mathrm{Var}(\mathrm{HT}(y))$. Further, we replace D2 by the fourth-order moment condition

- D4. The study variables $\{y_k\}_{k \in U}$ satisfy
$$\limsup_{N \to \infty} N^{-1} \sum_{k \in U} y_k^4 < \infty.$$

It can then be shown that $n\mathrm{E}[\{\widehat{\mathrm{V}}(N^{-1}\,\mathrm{HT}(y)) - \mathrm{Var}(N^{-1}\,\mathrm{HT}(y))\}^2] \to 0$, using bounding arguments analogous to those in (6), again yielding mean square consistency and design consistency. Taken together with the assumed asymptotic normality, normal confidence intervals can be readily constructed and will have the correct coverage in moderate to large samples, because

$$\big\{\widehat{V}\big(\mathrm{HT}(y)\big)\big\}^{-1/2}\big\{\mathrm{HT}(y) - t_y\big\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

The conditions D1–D4 are similar to those used by Breidt and Opsomer (2000) and are discussed further there. The motivation for the design assumptions D1 and D3 is that the sequence of sampling design behaves

similarly to simple random sampling (without replacement), in the sense that the moments of the sample membership indicators of $p_N(\cdot)$ have the same order as those of simple random sampling. In Breidt and Opsomer (2008), D1 was relaxed to D1*, as follows:

- D1*. For all N, $\min_{k\in U}\pi_k \geq \pi_N^* > 0$ where $N\pi_N^* \to \infty$, and there exists $\kappa \geq 0$ such that $N^{1/2+\kappa}(\pi_N^*)^2 \to \infty$ and
$$\max_{k\in U}\sum_{\ell\in U:\ell\neq k}\Delta_{k\ell}^2 = O(N^{-2\kappa})$$
as $N\to\infty$.

In this version, the factor κ allows a more explicit trade-off between how small one-way inclusion probabilities are allowed to be and how large the covariance between sample membership indicators, while still maintaining the design consistency of the resulting estimator. Other versions of these regularity conditions are possible and appear in the literature. To obtain consistency of the variance estimator under more general conditions, D3 could similarly be relaxed, but we do not pursue this further here.

So far, the discussion in this section concerned the asymptotic properties of the Horvitz–Thompson estimator. But the argument and framework used here apply more broadly to general survey estimators, including those that are not unbiased nor have an exact design variance expression as in (2). For those estimators (i.e., the majority of estimators used in practice), the large-sample properties of design consistency and asymptotic distribution are the primary statistical properties of interest. Before we discuss this further below, we note that the asymptotic results so far directly carry over to the difference estimator, since the difference estimator is just a shifted version of a Horvitz–Thompson estimator by (4) for any (fixed) $m(\cdot)$. This will be key in obtaining the asymptotic properties of model-assisted estimators.

4. MODEL-ASSISTED ESTIMATION

The difference estimator requires a method $m(\cdot)$ independent of the sample, but in practice it will commonly be the case that we use the sample data to build the predictive method. Model-assisted survey estimation approaches this problem by introducing a "working model" for purposes of prediction. Many such models can be written as

$$y_k = \mu(\boldsymbol{x}_k) + \varepsilon_k,$$

with random, zero-mean $\{\varepsilon_k\}$, so that $\{y_k\}_{k\in U}$ in the finite population are now modeled as realizations from a stochastic *superpopulation* model. Importantly, we will not require that this model be correct for the population, but for it to be useful, it should still contain some predictive power with respect to the survey variables of interest.

A general "recipe" for estimation and inference using auxiliary information proceeds as follows:

- If $\{\boldsymbol{x}_k, y_k)\}_{k\in U}$ were observed for the entire population, a standard statistical method to estimate $\mu(\cdot)$ would result in $m_N(\cdot)$, which depends on the population but is independent of the sample.
- Since only a sample is observed, estimate $m_N(\cdot)$ by $\widehat{m}(\cdot)$, which is not independent of the sample.
- Plug $\widehat{m}(\cdot)$ into the difference estimator form (4) to yield the *model-assisted estimator*:
$$\text{DIFF}(y,\widehat{m}) = \sum_{k\in U}\widehat{m}(\boldsymbol{x}_k) + \sum_{k\in s}\frac{y_k-\widehat{m}(\boldsymbol{x}_k)}{\pi_k}. \tag{7}$$
- For inference, assume that the estimator is approximately normally distributed for large samples and estimate the variance by
$$\widehat{V}\big(\text{DIFF}(y,\widehat{m})\big) = \sum_{k,\ell\in U}\Delta_{k\ell}\frac{y_k-\widehat{m}(\boldsymbol{x}_k)}{\pi_k}\frac{y_\ell-\widehat{m}(\boldsymbol{x}_\ell)}{\pi_\ell}\frac{I_kI_\ell}{\pi_{k\ell}}. \tag{8}$$

The above "recipe" is surprisingly flexible and broadly applicable, allowing for many different formulations of the working model and different estimation methods, as will be discussed further in later sections. We first describe in general terms how to obtain the asymptotic properties of the model-assisted estimator.

5. SURVEY ASYMPTOTICS II

It is useful to rewrite (7) as

$$\begin{aligned}
&\text{DIFF}(y,\widehat{m})\\
&\quad= \sum_{k\in U}\widehat{m}(\boldsymbol{x}_k) + \sum_{k\in s}\frac{y_k-\widehat{m}(\boldsymbol{x_k})}{\pi_k}\\
&\quad= \sum_{k\in U}m_N(\boldsymbol{x}_k) + \sum_{k\in U}\frac{(y_k-m_N(\boldsymbol{x}_k))I_k}{\pi_k} \qquad (9)\\
&\qquad+ \sum_{k\in U}(\widehat{m}(\boldsymbol{x}_k)-m_N(\boldsymbol{x}_k))\left(1-\frac{I_k}{\pi_k}\right)\\
&\quad= \text{DIFF}(y,m_N) + (\text{remainder}).
\end{aligned}$$

In other words, the model-assisted estimator is equal to the (infeasible, but exactly unbiased) difference estimator based on the population-level fit $m_N(\cdot)$, plus a

remainder term. The technical challenge for a specific model-assisted estimation scenario is then to show that the remainder in (9) is negligible relative to the difference estimator itself. How to do this depends on the estimation method, with smoothness conditions and Taylor approximations as the most common approach. But once that remainder is determined to be negligible, the model-assisted estimator immediately inherits the asymptotic properties of the corresponding difference estimator $\mathrm{DIFF}(y, m_N)$.

The model-assisted estimator will be asymptotically unbiased since the difference estimator is unbiased for any $m_N(\cdot)$, regardless of the working model $\mu(\cdot)$. Similarly, it will be design consistent and asymptotically normally distributed under similar conditions as were needed for the Horvitz–Thompson estimator. The variance of this asymptotic distribution will be equal to the variance of the corresponding difference estimator, that is,

$$\sum_{k,\ell\in U} \Delta_{k\ell} \frac{y_k - m_N(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m_N(\boldsymbol{x}_\ell)}{\pi_\ell},$$

which can be consistently estimated by the "plug-in" estimator in (8) under further mild conditions, regardless of quality of the working model $\mu(\cdot)$. And critically from a practical perspective, this asymptotic variance will be smaller than that of $\mathrm{HT}(y)$ provided the residuals $\{y_k - m_N(\boldsymbol{x}_k)\}$ have less variation than the raw values $\{y_k\}$, a reasonable expectation for predictive methods.

We now consider special cases of this general formulation of the model-assisted estimator.

6. GENERALIZED REGRESSION ESTIMATION

The best-known class of model-assisted survey estimators are *generalized regression* or GREG estimators (Cassel, Särndal and Wretman, 1976, Särndal, Swensson and Wretman, 1992, Chapter 6), generated from a working model of heteroskedastic multiple regression:

$$\begin{aligned} y_k &= \mu(\boldsymbol{x}_k) + \varepsilon_k = \boldsymbol{x}_k'\boldsymbol{\beta} + \varepsilon_k, \\ &\{\varepsilon_k\} \text{ uncorrelated}\big(0, \sigma_k^2\big). \end{aligned} \tag{10}$$

If the entire population were observed, the parameters $\boldsymbol{\beta}$ would be estimated via weighted least squares, with corresponding predictors given by

$$\begin{aligned} m_N(\boldsymbol{x}_k) &= \boldsymbol{x}_k'\boldsymbol{B}_N \\ &= \boldsymbol{x}_k'\Big(\sum_{j\in U} \frac{\boldsymbol{x}_j\boldsymbol{x}_j'}{\sigma_j^2}\Big)^{-1} \sum_{j\in U} \frac{\boldsymbol{x}_j y_j}{\sigma_j^2}. \end{aligned} \tag{11}$$

Since only a sample is observed, we estimate the finite population fit by plugging in HT estimators for the finite population totals in (11), yielding

$$\widehat{m}(\boldsymbol{x}_k) = \boldsymbol{x}_k'\widehat{\boldsymbol{B}} = \boldsymbol{x}_k'\Big(\sum_{j\in s} \frac{\boldsymbol{x}_j\boldsymbol{x}_j'}{\pi_j\sigma_j^2}\Big)^{-1} \sum_{j\in s} \frac{\boldsymbol{x}_j y_j}{\pi_j\sigma_j^2}. \tag{12}$$

Finally, plugging (12) into the model-assisted estimator (7), we have the GREG:

$$\mathrm{DIFF}\big(y, \boldsymbol{x}_k'\widehat{\boldsymbol{B}}\big) = \sum_{k\in U} \boldsymbol{x}_k'\widehat{\boldsymbol{B}} + \sum_{k\in s} \frac{y_k - \boldsymbol{x}_k'\widehat{\boldsymbol{B}}}{\pi_k}. \tag{13}$$

Särndal, Swensson and Wretman (1992) is a comprehensive treatment of such models, including their extensions to multiple stages or multiple phases of sampling. Well-known GREG examples (see, e.g., Cochran, 1977, Chapters 5–7) include the post-stratification estimator, in which $\boldsymbol{x}_k$ is a vector of indicators for the levels of a categorical covariate; the classical survey ratio estimator, in which x_k is a scalar, and the model is heteroskedastic regression through the origin; and the classical survey regression estimator, in which x_k is a scalar, and the model is homoskedastic simple linear regression.

To study the properties of GREG, we first follow (9) and write

$$\begin{aligned} \mathrm{DIFF}\big(y, \boldsymbol{x}_k'\widehat{\boldsymbol{B}}\big) &= \sum_{k\in U} \boldsymbol{x}_k'\widehat{\boldsymbol{B}} + \sum_{k\in s} \frac{y_k - \boldsymbol{x}_k'\widehat{\boldsymbol{B}}}{\pi_k} \\ &= \sum_{k\in U} \boldsymbol{x}_k'\boldsymbol{B}_N + \sum_{k\in U} \frac{(y_k - \boldsymbol{x}_k'\boldsymbol{B}_N)I_k}{\pi_k} \\ &\quad + (\widehat{\boldsymbol{B}}_N - \boldsymbol{B}_N)'\big(t_{\boldsymbol{x}} - \mathrm{HT}(\boldsymbol{x})\big) \\ &= \mathrm{DIFF}\big(y, \boldsymbol{x}_k'\boldsymbol{B}_N\big) + (\text{remainder}). \end{aligned}$$

The remainder is

$$\begin{aligned} &(\widehat{\boldsymbol{B}}_N - \boldsymbol{B}_N)'\big(t_{\boldsymbol{x}} - \mathrm{HT}(\boldsymbol{x})\big) \\ &\quad = \sum_{i=1}^{p} (\widehat{B}_i - B_{N,i})\big(t_{x_i} - \mathrm{HT}(x_i)\big), \end{aligned}$$

where each of the differences in the regression coefficients (sample estimate minus finite population parameter) is $O_p((N\pi_N^*)^{-1/2})$, as a smooth function of HT estimators, and each of the differences in the x_i-totals (finite population total minus HT estimator) is $O_p(N(N\pi_N^*)^{-1/2})$. Assuming that these rates can be applied uniformly across i, the overall rate for the remainder term is

$$O_p\big(N\big(N\pi_N^*\big)^{-1}\big) = o_p\big(N\big(N\pi_N^*\big)^{-1/2}\big)$$

as $N\pi_N^* \to \infty$. Hence, GREG behaves asymptotically like a difference estimator (see, e.g., Isaki and Fuller, 1982, Robinson and Särndal, 1983). It is asymptotically unbiased (and mean square consistent), regardless of the quality of the heteroskedastic regression model. Its variance is asymptotically equivalent to

$$\begin{aligned}&\operatorname{Var}\Bigg(\sum_{k\in U}\boldsymbol{x}_k'\boldsymbol{B}_N+\sum_{k\in U}(y_k-\boldsymbol{x}_k'\boldsymbol{B}_N)\frac{I_k}{\pi_k}\Bigg)\\&\quad=\sum_{k,\ell\in U}\Delta_{k\ell}\frac{y_k-\boldsymbol{x}_k'\boldsymbol{B}_N}{\pi_k}\frac{y_\ell-\boldsymbol{x}_k'\boldsymbol{B}_N}{\pi_\ell},\end{aligned}$$

and this asymptotic variance is smaller than that of $\mathrm{HT}(y)$ provided that the finite population regression residuals $\{y_k-\boldsymbol{x}_k'\boldsymbol{B}_N\}$ have less variation than the raw values $\{y_k\}$. The asymptotic variance can be estimated using the "plug-in" model-assisted variance estimator (8).

6.1 Weighting and Calibration for the GREG

A useful property of GREG is that it can also be written in weighted form:

$$\begin{aligned}&\mathrm{DIFF}\big(y,\boldsymbol{x}_k'\widehat{\boldsymbol{B}}\big)\\&\quad=\sum_{k\in s}\frac{y_k-\boldsymbol{x}_k'\widehat{\boldsymbol{B}}}{\pi_k}+\sum_{k\in U}\boldsymbol{x}_k'\widehat{\boldsymbol{B}}\\&\quad=\sum_{k\in s}\Bigg\{\frac{1}{\pi_k}+\big(t_{\boldsymbol{x}}-\mathrm{HT}(\boldsymbol{x})\big)'\Bigg(\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\pi_k}\Bigg)^{-1}\frac{\boldsymbol{x}_k}{\pi_k}\Bigg\}y_k\\&\quad=\sum_{k\in s}\omega_{ks}y_k.\end{aligned}$$

The GREG weights $\{\omega_{ks}\}$ have the form of the original HT weights $\{\pi_k^{-1}\}$ modified to take into account the $\boldsymbol{x}$-information. Importantly, the GREG weights depend on $\{\boldsymbol{x}_k\}_{k\in s}$, but not on $\{\boldsymbol{x}_k\}_{k\in U\setminus s}$ except through the known totals $t_{\boldsymbol{x}}$.

The GREG weights do not depend on y and can be applied to any response variable. This is useful in a multi-purpose survey, where many different responses y are collected, and a single set of weights can be applied to all of these responses. In particular, the GREG weights can be applied to the sampled $\boldsymbol{x}$ variables,

$$\begin{aligned}&\mathrm{DIFF}\big(\boldsymbol{x}',\boldsymbol{x}_k'\widehat{\boldsymbol{B}}\big)\\&\quad=\sum_{k\in s}\omega_{ks}\boldsymbol{x}_k'\\&\quad=\sum_{k\in s}\Bigg\{\frac{1}{\pi_k}+\big(t_{\boldsymbol{x}}-\mathrm{HT}(\boldsymbol{x})\big)'\Bigg(\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\pi_k}\Bigg)^{-1}\frac{\boldsymbol{x}_k}{\pi_k}\Bigg\}\boldsymbol{x}_k'\\&\quad=\mathrm{HT}\big(\boldsymbol{x}'\big)+\big(t_{\boldsymbol{x}}-\mathrm{HT}(\boldsymbol{x})\big)'\Bigg(\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\pi_k}\Bigg)^{-1}\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\pi_k}\\&\quad=t_{\boldsymbol{x}}'.\end{aligned}$$

In this case, we say that the weights $\{\omega_{ks}\}$ are *calibrated* to the $\boldsymbol{x}$-totals in the sense that the weighted sample estimates equal the known finite population totals. This calibration can be used by a survey agency to insure internal consistency across a statistical system. Clearly, GREG will be very efficient if y_k is approximately a linear combination of $\boldsymbol{x}_k$.

6.2 Recent Extensions of GREG

Cardot and Josserand (2011) discuss Horvitz–Thompson estimation for functional data, in which scalar observations y_k are replaced by functions $y_k(t)$. By generalizing the asymptotic framework described in Section 3 to the functional data setting, they obtain similar design-based properties as discussed earlier for $\mathrm{HT}(y)$. Cardot, Goga and Lardin (2013) propose a model-assisted estimator for functional data based on the following working model:

$$y_k(t)=\boldsymbol{x}_k'\boldsymbol{\beta}(t)+\varepsilon_k(t),$$

where $\boldsymbol{\beta}(t)$ is a vector of functional regression coefficients and the $\varepsilon_k(t)$ are independent stochastic processes. Applying the model-assisted recipe from Section 4, they show that the resulting functional regression estimator is asymptotically equivalent to a functional difference estimator, after which the usual asymptotic properties once again follow.

Beaumont, Haziza and Ruiz-Gazen (2013) propose a modified version of the Horvitz–Thompson estimator that attempts to remove the effect of influential points, which are defined in this design-based context as points with large conditional bias $\mathrm{CB}_k=\mathrm{E}[\mathrm{HT}(y)|I_k=1]-t_y$. The value of CB_k depends on y_k as well as on the design. The authors use a Huber function approach to downweight the contribution of points with values of $|\mathrm{CB}_k|$ above a threshold value. In the model-assisted extension of this robust estimator, by appealing to the asymptotic equivalence between GREG and DIFF estimators, they show that the conditional bias CB_k now depends on the residual $y_k-m_N(\boldsymbol{x}_k)$ instead of on the magnitude of the y_k themselves. The model-assisted estimator then mimics the robust Horvitz–Thompson estimator, by downweighting observations with large values of the model residual-based $|\mathrm{CB}_k|$.

6.3 Weaknesses of GREG

In practice, the GREG weight adjustments may be large, so that extreme weights and negative weights are possible. The effects of extreme weights are particularly noticeable in domain estimation, where the effect of outlying weights is not diluted by large sample size. Many survey estimation procedures have been motivated by these weaknesses of GREG. In this review, we emphasize "regression-like" methods, many of which attempt to trim, smooth or otherwise stabilize the weights through model specifications other than the heteroskedastic regression. Among these are methods based on linear mixed models, nonparametric or semiparametric regression, or other flexible prediction techniques that can broadly be described as statistical or machine learning.

7. LINEAR MIXED MODELS FOR MODEL-ASSISTED ESTIMATION

Let z_k be a $K \times 1$ vector of known covariates, in addition to the $p \times 1$ vector $\boldsymbol{x}_k$. Write $\boldsymbol{X}_U = [\boldsymbol{x}_k']_{k\in U}$, $\boldsymbol{Z}_U = [z_k']_{k\in U}$, and $\boldsymbol{C}_U = [\boldsymbol{X}_U, \boldsymbol{Z}_U]$. Consider the following linear mixed model as a working model:

$$\boldsymbol{y}_U = \boldsymbol{X}_U\boldsymbol{\beta} + \boldsymbol{Z}_U\boldsymbol{b} + \boldsymbol{\varepsilon}_U \tag{14}$$

where

$$\mathrm{E}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\varepsilon}_U\end{bmatrix} = \boldsymbol{0}, \quad \mathrm{Var}\left(\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\varepsilon}_U\end{bmatrix}\right) = \sigma^2\begin{bmatrix}\lambda^{-2}\boldsymbol{Q} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix},$$

with $\boldsymbol{Q}$ known and positive definite, σ^2 unknown, and λ fixed in advance. This is of course an incredibly broad and useful model class, allowing various continuous and categorical covariates $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ and a broad range of variance–covariance structures through the choices of $\boldsymbol{Q}$ and $\boldsymbol{R}$. The model class includes multiple regression, analysis of variance models, and many longitudinal models, along with penalized splines, varying-coefficient models, semiparametric additive models, low-rank kriging, and many more. See, for example, Ruppert, Wand and Carroll (2003).

Linear mixed models are widely used in estimation for complex surveys, particularly in small area estimation (Fay and Herriot, 1979, Battese, Harter and Fuller, 1988, Datta and Ghosh, 1991, Ghosh and Rao, 1994, Rao, 2003, Opsomer et al., 2008). They have also been used to stabilize weights or relax constraints in calibration estimation and related methods, including robust case weighting (Bardsley and Chambers, 1984, Chambers, 1996), ridge calibration (Rao and Singh, 1997, Beaumont and Bocci, 2008, Montanari and Ranalli, 2009), and other methods for satisfying range restrictions (Park and Fuller, 2005) or smoothing weights (Lazzeroni and Little, 1998, Elliott and Little, 2000). Here, we focus on the use of linear mixed models for extensions of GREG (Fuller, 2002, Zheng and Little, 2003, 2004, Breidt, Claeskens and Opsomer, 2005, Park and Fuller, 2009, Guggemos and Tillé, 2010), returning to the fitting of linear mixed models as a regularization/penalization method in Section 8.3.

If the entire finite population were observed, the best linear unbiased estimators (BLUEs) of $\boldsymbol{\beta}$ and best linear unbiased predictors (BLUPs) of $\boldsymbol{b}$ would be obtained via

$$\boldsymbol{B}_N = (\boldsymbol{C}_U'\boldsymbol{C}_U + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}_U'\boldsymbol{y}_U,$$

where $\boldsymbol{\Lambda} = \mathrm{blockdiag}(\boldsymbol{0}, \lambda^2\boldsymbol{Q}^{-1})$. The corresponding BLUPs of individual y_k values would then be

$$m_N(\boldsymbol{c}_k) = \boldsymbol{c}_k'\boldsymbol{B}_N = \boldsymbol{c}_k'(\boldsymbol{C}_U'\boldsymbol{C}_U + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}_U'\boldsymbol{y}_U$$

with $\boldsymbol{c}_k' = (\boldsymbol{x}_k', z_k')$. The corresponding survey-weighted versions are

$$\widehat{\boldsymbol{B}} = \left(\sum_{k\in s}\frac{\boldsymbol{c}_k\boldsymbol{c}_k'}{\pi_k} + \boldsymbol{\Lambda}\right)^{-1}\sum_{k\in s}\frac{\boldsymbol{c}_k y_k}{\pi_k},$$

and $\widehat{m}(\boldsymbol{c}_k) = \boldsymbol{c}_k'\widehat{\boldsymbol{B}}$. Plugging these in to the difference estimator, we have the linear mixed model estimator

$$\mathrm{DIFF}\big(y, \boldsymbol{c}_k'\widehat{\boldsymbol{B}}\big) = \sum_{k\in U}\boldsymbol{c}_k'\widehat{\boldsymbol{B}} + \sum_{k\in s}\frac{y_k - \boldsymbol{c}_k'\widehat{\boldsymbol{B}}}{\pi_k}. \tag{15}$$

As this form shows, the general recipe from Section 4 continues to apply here. Hence, the asymptotic variance can again be obtained and variance estimation will be based on (8).

Like the GREG, the linear mixed model estimator can be written in weighted form:

$$\begin{aligned}
&\mathrm{DIFF}\big(y, \boldsymbol{c}_k'\widehat{\boldsymbol{B}}\big)\\
&\quad= \sum_{k\in s}\frac{y_k - \boldsymbol{c}_k'\widehat{\boldsymbol{B}}}{\pi_k} + \sum_{k\in U}\boldsymbol{c}_k'\widehat{\boldsymbol{B}}\\
&\quad= \sum_{k\in s}\left\{\frac{1}{\pi_k} + (t_{\boldsymbol{c}} - \mathrm{HT}(\boldsymbol{c}))'\left(\sum_{k\in s}\frac{\boldsymbol{c}_k\boldsymbol{c}_k'}{\pi_k} + \boldsymbol{\Lambda}\right)^{-1}\frac{\boldsymbol{c}_k}{\pi_k}\right\}y_k\\
&\quad= \sum_{k\in s}\omega_{ks}y_k.
\end{aligned}$$

The weights $\{\omega_{ks}\}$ are calibrated to the $\boldsymbol{x}$-population totals, $\mathrm{DIFF}(\boldsymbol{x}', \boldsymbol{c}_k'\widehat{\boldsymbol{B}}) = t_{\boldsymbol{x}}'$, but are not calibrated to the z-population totals, $\mathrm{DIFF}(z', \boldsymbol{c}_k'\widehat{\boldsymbol{B}}) \neq t_z'$. This relaxation of the calibration constraints provides useful flexibility for GREG-type estimators.

8. MODEL-ASSISTED METHODS BASED ON STATISTICAL LEARNING TECHNIQUES

8.1 Model Calibration Approach

Outside of the linear model, many procedures lead to predictions that are not linear combinations of the observed data, thus complicating the calibration and weighting properties of GREG-type estimators. Wu and Sitter (2001) propose a simple and effective method for incorporating nonlinear predictions into model-assisted estimators, by using an arbitrary predictor $\widehat{m}(\boldsymbol{x}_k)$ as the covariate z_k in the working model:

$$y_k = \beta z_k + \varepsilon_k, \quad \{\varepsilon_k\} \text{ uncorrelated}(0, \sigma^2). \tag{16}$$

Estimating β by $B_N = \sum_{k\in U} z_k y_k (\sum_{k\in U} z_k^2)^{-1}$ at the finite population level, B_N by

$$\widehat{B} = \frac{\sum_{k\in s} z_k y_k/\pi_k}{\sum_{k\in s} z_k^2/\pi_k} = \frac{\sum_{k\in s} \widehat{m}(\boldsymbol{x}_k) y_k/\pi_k}{\sum_{k\in s} \widehat{m}(\boldsymbol{x}_k)^2/\pi_k}$$

at the sample level, and plugging in to the model-assisted estimator form then leads to the *model calibration estimator* of Wu and Sitter (2001):

$$\text{DIFF}\big(y, \widehat{B}\widehat{m}(\boldsymbol{x}_k)\big) = \sum_{k\in U} \widehat{B}\widehat{m}(\boldsymbol{x}_k) + \sum_{k\in s} \frac{y_k - \widehat{B}\widehat{m}(\boldsymbol{x}_k)}{\pi_k}.$$

Wu (2003) later showed that, under certain regularity conditions on the sampling design, the model calibration estimator is optimal in the sense that it minimizes the superpopulation model expectation of the asymptotic design variance over a class of calibration estimators.

8.2 Kernel Methods and Local Regression

Kernel methods assume that the working model is locally simple, like constant or linear, and globally smooth, then estimate the local regression function using only nearby points as determined by a kernel weighting function. In this section, we focus on local polynomial regression and its extensions in the model-assisted context.

Local polynomial regression. Breidt and Opsomer (2000) considered a working model in which $\mu(\cdot)$ is a smooth function of scalar x, to be estimated by *local polynomial regression*, in which the smooth function is approximated locally at x_k by a q-th order polynomial, fitted at the finite population level via weighted least squares regression with weights given by a kernel function centered at x_k:

$$\begin{aligned} m_N(x_k) &= (1, 0, \ldots, 0) \\ &\quad \cdot (\boldsymbol{X}_{Uk}' \boldsymbol{W}_{Uk} \boldsymbol{X}_{Uk})^{-1} \boldsymbol{X}_{Uk}' \boldsymbol{W}_{Uk} \boldsymbol{y}_U \end{aligned} \tag{17}$$

where

$$\boldsymbol{X}_{Uk} = \begin{bmatrix} 1 & x_j - x_k & \cdots & (x_j - x_k)^q \end{bmatrix}_{j\in U},$$

$$\boldsymbol{W}_{Uk} = \operatorname{diag}\left\{ \frac{1}{h} K\left(\frac{x_j - x_k}{h} \right) \right\}_{j\in U}$$

and $\boldsymbol{y}_U' = [y_1, y_2, \ldots, y_N]$. Then, letting

$$\boldsymbol{X}_{sk} = \begin{bmatrix} 1 & x_j - x_k & \cdots & (x_j - x_k)^q \end{bmatrix}_{j\in s},$$

$$\boldsymbol{W}_{sk} = \operatorname{diag}\left\{ \frac{1}{\pi_j h} K\left(\frac{x_j - x_k}{h} \right) \right\}_{j\in s}$$

and $\boldsymbol{y}_s' = [y_j]_{j\in s}$, we define

$$\widehat{m}_{\text{LPR}}(x_k) = (1, 0, \ldots, 0)(\boldsymbol{X}_{sk}' \boldsymbol{W}_{sk} \boldsymbol{X}_{sk})^{-1} \boldsymbol{X}_{sk}' \boldsymbol{W}_{sk} \boldsymbol{y}_s$$

as the design-weighted version of (17). The local polynomial survey regression estimator, LPR, is then

$$\begin{aligned} &\text{DIFF}\big(y, \widehat{m}_{\text{LPR}}(x_k)\big) \\ &\quad = \sum_{k\in U} \widehat{m}_{\text{LPR}}(x_k) + \sum_{k\in s} \frac{y_k - \widehat{m}_{\text{LPR}}(x_k)}{\pi_k} \\ &\quad = \sum_{k\in U} m_N(x_k) + \sum_{k\in U} \frac{(y_k - m_N(x_k)) I_k}{\pi_k} \\ &\qquad + \sum_{k\in U} \big(\widehat{m}_{\text{LPR}}(x_k) - m_N(x_k)\big)\left(1 - \frac{I_k}{\pi_k}\right) \\ &\quad = \text{DIFF}(y, m_N) + (\text{remainder}). \end{aligned}$$

Under a fixed-size design with inclusion probabilities bounded away from zero and sampling rate $n_N N^{-1} \to \pi > 0$, with kernel bandwidth $h_N \to 0$ and $N h_N^2/(\log\log N) \to \infty$, and with weak design dependence as measured by second-order through fourth-order inclusion probabilities, Breidt and Opsomer (2000), Lemma 5, show that the remainder is of order

$$o_p\big(N n_N^{-1/2}\big),$$

so that LPR again inherits the properties of the difference estimator, including its asymptotic variance and variance estimator. Further, they show that $\text{LPR}(y) = \sum_{k\in s} \omega_{ks} y_k$ with weights $\{\omega_{is}\}$ independent of y. For a qth-order local polynomial, these weights are calibrated to powers of x:

$$\sum_{k\in s} \omega_{ks} x_k^\ell = \sum_{k\in U} x_k^\ell \quad (\ell = 0, 1, \ldots, q).$$

Hence, LPR will be particularly effective if y is approximately a qth order polynomial in x. Unlike GREG, use of LPR requires availability of x_k for all $k \in U$.

Montanari and Ranalli (2005) added a model calibration step as in Wu and Sitter (2001) to LPR to gain additional calibration with respect to the working model. In simulations, the model-calibrated LPR performed slightly better than the original version.

Other extensions include Deville and Goga (2004), who applied LPR to improve the efficiency of survey estimators when samples are taken on two occasions, and Aragon, Goga and Ruiz-Gazen (2006), who considered quantile estimation. Rueda, Sánchez-Borrego and Arcos (2009) construct a jump-preserving model assisted estimator by adapting LPR to allow for discontinuities in $\mu(\cdot)$. They compare model-assisted and model-based versions of LPR and its jump-preserving counterparts via simulation, while Sánchez-Borrego, Rueda and Muñoz (2012) conduct an empirical study of these estimators using data on breast cancer prevalence in 40 European countries.

Additive models. Breidt et al. (2007) considered a model-assisted estimator in which the mean of the working model is a semiparametric additive model,

$$\mu(\boldsymbol{x}_k) = \mu_1(x_{1k}) + \cdots + \mu_q(x_{qk}) + \boldsymbol{x}_k'\boldsymbol{\beta}, \tag{18}$$

where the $\mu_1(\cdot), \ldots, \mu_q(\cdot)$ are unknown smooth functions of their respective scalar arguments. They provided a design-weighted backfitting algorithm for efficient estimation of the semiparametric additive model, alternating between local polynomial regression for estimation of the smooth functions, and design-weighted least squares for estimation of the regression coefficients $\boldsymbol{\beta}$. For the special case of $q = 1$, they proved design consistency, derived a consistent variance estimator, and established asymptotic normality under assumptions similar to Breidt and Opsomer (2000).

Model (18) is the special case (with identity link) of the generalized additive model

$$\mu(\boldsymbol{x}_k) = g\{\mu_1(\boldsymbol{x}_k) + \cdots + \mu_q(\boldsymbol{x}_k) + \boldsymbol{x}_k'\boldsymbol{\beta}\}, \tag{19}$$

where $g(\cdot)$ is a known link function and $\mu_1(\cdot), \ldots, \mu_q(\cdot)$ are unknown smooth functions, each operating on one or more components of $\boldsymbol{x}_k$. Opsomer et al. (2007) constructed a model-assisted estimator using the generalized additive model (19) as the mean of the working model, fitted the model using local scoring, and applied it to a forest inventory problem of estimating the total of a binary "forest/nonforest" variable. The approach of Wu and Sitter (2001) was used to generate weights to be applied to all the survey variables.

Wang and Wang (2011) focused on the nonparametric additive components of (18) and fitted the model using a two-stage spline-backfitted local polynomial regression, in which splines were used in each backfitting step for initial estimation and removal of all additive components except the one of interest, and then local polynomial regression was applied to estimate the smooth component. This two-step estimator is computationally efficient and allows for formal derivation of asymptotic properties of the model-assisted estimator for $q \geq 1$.

8.3 Splines

Some of the disadvantages of kernel-based methods like LPR include the difficulties of adapting the kernel to incorporate multiple covariates, especially combinations of categorical and continuous covariates, and the computational challenges for datasets with regions of sparse data. These disadvantages are largely overcome by using a large number of splines or other basis functions, together with selection or regularization/penalization to control the complexity of the model.

Penalized splines. In Breidt, Claeskens and Opsomer (2005), the working model is a smooth function of scalar x. A set of K fixed, known knots $\{\kappa_j\}_{j=1}^K$ partitions the range of x, and the working model is the linear mixed model:

$$\begin{aligned} y_k &= [1, x_k, \ldots, x_k^q]\boldsymbol{\beta} \\ &\quad + \left[(x_k - \kappa_1)_+^q \;\; (x_k - \kappa_2)_+^q \;\; \cdots \;\; (x_k - \kappa_K)_+^q\right]\boldsymbol{b} \\ &\quad + \varepsilon_k \\ &= \boldsymbol{c}_k' \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} + \varepsilon_k \end{aligned} \tag{20}$$

with $(z)_+ = \max\{0, z\}$,

$$\mathrm{E}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\varepsilon}_U \end{bmatrix} = \boldsymbol{0}, \quad \mathrm{Var}\left(\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\varepsilon}_U \end{bmatrix}\right) = \sigma^2 \begin{bmatrix} \lambda^{-1}\boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_n \end{bmatrix},$$

and λ chosen a priori to give specified degrees of freedom in the smooth. As $\lambda \to \infty$, the working model corresponds to a global qth-order polynomial, while as $\lambda \to 0$, the working model corresponds to a piecewise qth-order polynomial between the knots.

With this model specification, the penalized spline (p-spline) estimator of Breidt, Claeskens and Opsomer (2005) behaves like the difference estimator: under a standard asymptotic framework with K fixed, the p-spline estimator is mean square consistent for t_y; its variance is asymptotically equivalent to that of the difference estimator at the finite population fit; the standard plug-in variance estimator is consistent; and it

has smaller asymptotic variance than HT provided the residuals $\{y_k - m_N(\boldsymbol{c}_k)\}$ have less variation than the raw values $\{y_k\}$. McConville and Breidt (2013) extend these results to an asymptotic setting with $K \to \infty$.

As with the general linear mixed model estimator in Section 7, the p-spline estimator can be written in weighted form, with weights that are calibrated to the population totals $t_{\boldsymbol{x}} = [\sum_{k\in U} x_k^\ell]_{\ell=0}^q$, but are not calibrated to the totals of the basis functions, t_z. That is, the estimator is fitted using a potentially large set of basis functions, but only a small number of calibration constraints are enforced.

Like the local polynomial survey regression estimator, the p-spline model-assisted estimator has good robustness properties inherited from its difference estimator form. Both have comparable efficiency to GREG when a parametric working model is correct, but better efficiency when the parametric working model is incorrect. Both have better-behaved weights than GREG (e.g., almost never negative). Unlike local polynomial regression and other kernel methods, penalized spline regression is readily formulated as a linear mixed model, and thus can be conveniently extended to include additional $\boldsymbol{x}_k$ variables, continuous or categorical, and additional $\boldsymbol{z}_k$ variables, continuous or categorical. This extended "semiparametric" model will yield weights that are calibrated on the $\boldsymbol{x}_k$ variables but not calibrated on the $\boldsymbol{z}_k$ variables. Any continuous components of $\boldsymbol{x}_k$ can enter the model parametrically or nonparametrically, through additional spline terms. Breidt, Claeskens and Opsomer (2005) discuss extension of model (20) to the semiparametric additive model (18), which is straightforward under the p-spline formulation: simply add fixed effects for any additional parametric terms $\boldsymbol{x}_k'\boldsymbol{\beta}$ and random effects for the additional nonparametric terms $\mu_2(x_{2k}), \ldots, \mu_q(x_{qk})$. Additional random effects to describe correlation structure can also be added as with any linear mixed model.

Regression splines. Goga (2004, 2005) studies another class of nonparametric model-assisted estimators based on regression splines. In both papers, Goga uses unpenalized regression splines, dividing the domain by using K knots, constructing a B-spline basis for the set of knots, and letting $K \to \infty$ so that the B-splines become dense on the domain. Goga (2005) shows that the model-assisted regression spline estimator is asymptotically design-unbiased and consistent, proposes a design-based variance approximation, and shows that the anticipated variance is asymptotically equivalent to the Godambe–Joshi lower bound. Simulations show that the regression spline estimator has good properties. Goga (2004) constructs model-assisted estimators in the case of sampling on two occasions, with complete auxiliary information available on each occasion.

8.4 Neural Networks and Related Methods

We now turn to a class of prediction methods in which new covariates are derived as linear combinations of the original $\boldsymbol{x}_k$, and then the working model postulates that the mean response is a nonlinear function of the new covariates; see, for example, Hastie, Tibshirani and Friedman (2001), Chapter 11. Included in this class are neural network models and projection pursuit models.

Neural networks. Montanari and Ranalli (2005) developed a model-assisted estimator for a working model that is a feedforward neural network with skip-layer connections,

$$\mu(\boldsymbol{x}_k) = \boldsymbol{x}_k'\boldsymbol{\beta} + \sum_{j=1}^{M} \alpha_j a(\boldsymbol{\gamma}_j'\boldsymbol{x}_k), \tag{21}$$

where $a(\cdot)$ is a known *activation function*, often a sigmoidal function, and $\boldsymbol{\beta}$, $\alpha_1, \ldots, \alpha_M$ and $\{\boldsymbol{\gamma}_j\}_{j=1}^M$ are unknown parameters. They then used model calibration as in Wu and Sitter (2001) to generate weights. Montanari and Ranalli (2005) prove design consistency and asymptotic normality of the model-assisted estimator, and provide a consistent variance estimator. In simulations, the model-calibrated neural network estimator outperforms the local polynomial regression estimator.

Projection pursuit and single-index models. Projection pursuit (Friedman and Stuetzle, 1981) is closely related to neural network modeling, replacing the specification $\{\alpha_j a(\cdot)\}_{j=1}^M$ in (21) with unknown smooth functions $\{\alpha_j(\cdot)\}_{j=1}^M$,

$$\mu(\boldsymbol{x}_k) = \sum_{j=1}^{M} \alpha_j(\boldsymbol{\gamma}_j'\boldsymbol{x}_k), \tag{22}$$

where $\{\boldsymbol{\gamma}_j\}_{j=1}^M$ are unknown parameters. The unknown functions are estimated via some flexible smoothing method; see, for example, Hastie, Tibshirani and Friedman (2001), Chapter 11.2. While we are not aware of model-assisted survey estimators that use projection pursuit, Wang (2009) considered the special case of (22) with $M = 1$, also known as the *single-index model*. Rather than use

$$\mu(\boldsymbol{x}_k) = \alpha(\boldsymbol{\gamma}'\boldsymbol{x}_k), \tag{23}$$

directly, Wang (2009) used an approximation to $\mu(\cdot)$ given by

$$\mu_*(z) = \mathrm{E}\big[\mu(\boldsymbol{x}_k) \mid \boldsymbol{\gamma}'\boldsymbol{x}_k = z\big]$$

where the expectation is with respect to the model in (23), and estimated $\mu_*(\cdot)$ via polynomial splines. Under design assumptions like those of Breidt and Opsomer (2000), Wang (2009) proved design consistency and asymptotic normality of the model-assisted estimator, and provided a consistent variance estimator.

8.5 K-Nearest Neighbor Methods

The K-nearest neighbor method (KNN) predicts y_k by averaging "nearby" $\{y_\ell\}_{\ell\in L_k\cap s}$ in a covariate-determined neighborhood of element k. Specifically, the neighborhood L_k of size $|L_k| = K$ is formed by finding the K vectors $\{\boldsymbol{x}_\ell\}_{\ell\in s}$ that are closest to $\boldsymbol{x}_k$ under some metric, such as Euclidean distance. The predictor then has the form

$$\widehat{m}(\boldsymbol{x}_k) = \frac{1}{K}\sum_{\ell\in L_k} y_k;$$

weighted versions of the average are also possible. The working model for KNN is that $\mu(\boldsymbol{x})$ is well-approximated by a locally constant function (Hastie, Tibshirani and Friedman, 2001, Chapter 2.4). Such methods are now widely used in the international forest inventory community for combining ground-based measurements with remotely-sensed imagery; see McRoberts, Tomppo and Næsset (2010) and the references therein. Baffetta et al. (2009) developed a design-based model-assisted estimator using KNN, arguing intuitively (Baffetta et al., 2009, Section A.3) that the remainder term in (9) should be negligible. The authors evaluate the estimator via simulation and via application to estimation of timber volume in a forest inventory for the northeastern part of Tuscany, Italy. Covariates in the application were obtained as seven spectral bands from Landsat imagery. Extensions of the methodology are reported in Baffetta, Corona and Fattorini (2010).

8.6 Tree-Based Methods

Hastie, Tibshirani and Friedman (2001), Chapter 9.2, describe tree-based methods as those which partition the space of available covariates into a set of rectangles, then fit a simple model, like a constant, on each such rectangle.

Regression trees. While we are not aware of model-assisted estimators that take full advantage of recursive partitioning of samples for the construction of regression trees, like CART (Breiman et al., 1984) or MARS (Friedman, 1991), much of the asymptotic theory is already in place in the work of Toth and Eltinge (2011), who establish asymptotic design L^2 consistency of survey-weighted regression trees as estimators of quite general regression functions. Such survey-weighted regression trees could be readily incorporated into the model-assisted estimator (7). In modern practice, trees are often built repeatedly on randomly-selected subsets of the original data with randomly-selected covariates, then averaged, resulting in *random forests*; see Section 8.8 below.

Endogenous post-stratification. One example of a tree-based method was motivated (appropriately enough) by an application in forestry surveys. Recall that the post-stratification estimator (PSE) is a special case of GREG with indicators for categorical covariates:

$$\mathrm{DIFF}\big(y, \boldsymbol{x}'\widehat{\boldsymbol{B}}\big) = \sum_{k\in U}\boldsymbol{x}_k'\widehat{\boldsymbol{B}} + \sum_{k\in s}\frac{y_k - \boldsymbol{x}_k'\widehat{\boldsymbol{B}}}{\pi_k}$$

where $\boldsymbol{x}_k = [\mathbf{1}_{\{k\in U_h\}}]_{h=1}^H$ and $U = \bigcup_{h=1}^H U_h$ is a partition of the population. In the PSE, the indicators are known for the sample and their sums $\sum_{k\in U}\mathbf{1}_{\{k\in U_h\}} = |U_h| = N_h$ are known for the population. The working model for the PSE is a constant mean within post-strata.

In forestry surveys, it is desirable to construct post-strata based on classification of satellite imagery. Supervised classification of images requires ground truth data, which are available from the survey. Hence, the survey data train the classification algorithm and the classified image post-stratifies the survey data, a cycle that Breidt and Opsomer (2008) refer to as *endogenous post-stratification*.

Expressing the image classification as a classification of the predictions from a fitted model, we have

$$\widetilde{\boldsymbol{x}}_k = [\mathbf{1}_{\{\tau_{h-1}<\widehat{m}(\boldsymbol{x}_k)\le\tau_h\}}]_{h=1}^H,$$

where the $\{\tau_h\}_{h=0}^H$ are known break points, and the endogenous post-stratification estimator (EPSE) is

$$\mathrm{DIFF}\big(y, \widetilde{\boldsymbol{x}}'\widehat{\boldsymbol{B}}\big) = \sum_{k\in U}\widetilde{\boldsymbol{x}}_k'\widehat{\boldsymbol{B}} + \sum_{k\in s}\frac{y_k - \widetilde{\boldsymbol{x}}_k'\widehat{\boldsymbol{B}}}{\pi_k}.$$

Because this is a kind of partition of covariate space into rectangles followed by piecewise constant model fitting, it can be understood as a tree-based method. Breidt and Opsomer (2008) showed that EPSE is asymptotically equivalent to PSE for smooth parametric models, while Dahlke et al. (2013) extended these

results to a wide class of nonparametric $m_N(\cdot)$. Tipton, Opsomer and Moisen (2013) used linear regression, spline regression and random forests in EPSE, and investigated the effect of estimation and optimal selection of unknown break points $\{\tau_h\}_{h=0}^{H}$. See, for example, McRoberts, Næsset and Gobakken (2013), Næsset et al. (2013) and the references therein for various applications of the EPSE methodology.

8.7 Model Selection and Shrinkage Methods

In some applications, such as natural resource surveys, the dimension of the covariate vector $\boldsymbol{x}_k$ is large. For example, in forest inventory, it is common that $\boldsymbol{x}_k$ includes "wall-to-wall" remote sensing, like satellite imagery or high-altitude aerial photography; information like elevation, aspect and slope from digital elevation models; and other data products derived in a geographic information system. Some of this auxiliary information may be highly correlated and some may have poor predictive ability for response variables and, therefore, model selection and/or coefficient shrinkage are appropriate to improve the efficiency of survey regression estimators of finite population totals, and also to stabilize the weights.

Silva and Skinner (1997) considered the use of best subsets and forward stepwise regression to estimate a finite population quantity under simple random sampling without replacement, effectively setting some coefficients to zero. Shrinkage methods that move coefficients toward zero can be very effective in improving prediction accuracy and in improving the properties of the GREG weights. Such methods include ridge calibration (Rao and Singh, 1997, Beaumont and Bocci, 2008, Montanari and Ranalli, 2009); see Section 7 above for other methods for satisfying range restrictions.

The "least absolute shrinkage and selection operator" (lasso) method proposed by Tibshirani (1996) simultaneously performs model selection and regularized coefficient estimation by shrinking coefficients to zero. The lasso method finds coefficients which minimize the sum of the squared residuals subject to a constraint on the sum of the absolute value of the coefficients.

McConville (2011) studied a model-assisted survey regression estimator in which survey-weighted lasso regression coefficients

$$\widehat{\boldsymbol{B}}^{(L)} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta})^T \boldsymbol{\Pi}_s^{-1} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}) + \lambda_N \sum_{j=1}^{p} |\beta_j|$$

are plugged into the GREG. For a sequence of finite populations and probability sampling designs, McConville (2011) derives asymptotic properties of the lasso survey regression estimator, including design consistency and central limit theory for the estimator and design consistency of a variance estimator. Further, lasso survey regression weights are developed, using either a model calibration approach as in Wu and Sitter (2001) or a ridge regression approximation following Tibshirani (1996). The results are extended to allow use of the adaptive lasso (Zou, 2006). In simulation studies with many highly-correlated covariates, lasso has much smaller weight variation than GREG (and essentially no negative weights), and has much lower mean squared error than GREG, particularly for domains.

8.8 Ensemble Learning and Related Methods

There are now many ensemble methods available in the statistical and machine learning literature. For purposes of this discussion, an ensemble method constructs not one but M working predictors, $\widehat{\mu}_j(\cdot)$, using some subset of the observed data and some (quite general) prediction method for the jth predictor. The ensemble method then uses the weighted average

$$\widehat{\mu}(\boldsymbol{x}_k) = \sum_{j=1}^{M} \omega_j \widehat{\mu}_j(\boldsymbol{x}_k) \tag{24}$$

for prediction. Key differences among ensemble predictors are typically determined by the choice of the weights, $\{\omega_j\}$. These ensembles can combine the predictions from many "weak" predictors into a single, more powerful predictor.

Model averaging. Li and Opsomer (2006) considered model averaging as a way to avoid having to select the "best" model in constructing a regression estimator, which can be difficult in settings with many covariates and/or which require selecting values for tuning parameters. The idea is that if a set of M estimators are all (approximately) unbiased, then a linear combination of these estimators will continue to be unbiased but can often have lower variance. Li and Opsomer (2006) conducted a simulation study comparing a number of simple model averaging approaches and generally found modest improvements in efficiency across different scenarios. This topic is still mostly unexplored and might become more important as larger quantities of auxiliary information become available in the future.

Random forests. Tipton, Opsomer and Moisen (2013) uses random forests (Breiman, 2001) in an endogenous post-stratification (that is, embedded within another tree-based method), but we are not aware of direct uses of random forests in a model-assisted survey estimator.

Bagging. Wang, Opsomer and Wang (2014) consider bootstrap aggregating or bagging (Breiman, 1996) in the design-based setting, showing the design consistency of bagged survey estimators and demonstrating the improved efficiency of the bagged estimators via simulation. They also show how to construct bagged estimators using replicate survey weights, which are often provided with survey data to allow for variance estimation. While we are not aware of bagged model-assisted survey estimators, this might be an interesting area for future research.

9. DISCUSSION

As we hope this overview of modern approaches to model-assisted estimation has illustrated, the range of methods that can be applied to improve design-based survey estimators has dramatically increased in the last two decades. We expect it to continue to do so, as survey statisticians continue to take advantage of new methods being developed in other areas of statistics. These new methods address a pressing need of statistical agencies conducting surveys, because of rising costs, increasing demands for more precise estimates at smaller scales and a general desire to maximally use known information about the target population in the survey estimates.

At the same time as these new methods are being developed, a shift in thinking about model-assisted estimation seems to be occurring. While this is a bit of an over-simplification, the traditional focus was on adjusting estimators to match interpretable population control totals. This can be clearly seen in post-stratification and ratio estimation, but is also present to a large extent in GREG approaches in general. In contrast, the new focus is on prediction, with methods being evaluated on their ability to generate model predictions that lead to precise model-assisted estimators. This view of statistical methods as primarily prediction tools is related to the ideas of *statistical learning* (Hastie, Tibshirani and Friedman, 2001), and many of the methods described in this article are familiar to researchers and practitioners in that area.

While good predictive ability is a key consideration in assessing the suitability of a statistical learning tool in the model-assisted context, other considerations are whether it can incorporate design weights and whether it produces good predictions across a range of response variables. The former consideration is important to maintain the design consistency of the resulting estimator. The latter concerns the fact that the main mode of application of model-assisted estimation is through the creation of survey weights, which are applied to all the variables in the survey. Hence, a method that leads to highly precise estimates for some variables but imprecise estimates for others is not as suitable as one that leads to good precision across all variables. This depends on the interplay between the survey variables, auxiliary variables and the methods themselves, so it is difficult to make general statements about which methods are better than others in which contexts. Nevertheless, practical applications of these methods will entail an evaluation of their robustness across survey variables and target estimates. Weight stability is often considered in this context, because highly variable weights can lead to imprecise estimates for variables that are poorly or negatively correlated with the weights.

Survey data are often used not only for estimation of finite population quantities, but also for model fitting (e.g., Lumley and Scott, same issue). In this context, parameter estimates are typically obtained by solving sample-weighted versions of finite-population estimating equations (Binder, 1983). For example, *pseudo-maximum likelihood estimators (PMLE)* are obtained by setting the survey-weighted score vector equal to zero and solving for the unknown parameters. The survey weights in PMLE and related estimators could be replaced by any of the model-assisted weights described in this article, which would have an impact on properties of the resulting parameter estimators. See Lumley and Scott (Section 3, same issue) and the references therein for an introduction to the use of *influence functions* for studying the effect of model-assisted weights on parameter estimation, a relatively unexplored area.

One aspect of model-assisted estimation that we hoped to convey is that across the range of methods discussed here, the underlying principles for the construction of estimators and the study of their statistical properties are surprisingly simple. The construction recipe in Section 4 enables practitioners to incorporate covariates and prediction methods from a wide range of sources, and the asymptotic framework of Section 3 and Section 5 shows a general path of statistical evaluation and the development of inference tools, with the

difference estimator based on the population fit of the prediction method forming the crucial link between the design-based and the model-based components.

REFERENCES

ARAGON, Y., GOGA, C. and RUIZ-GAZEN, A. (2006). Estimation non-paramétrique de quantiles en présence d'information auxiliaire. In *Méthodes D'Enquêtes et Sondages. Pratiques Européenne et Nord-américaine* (P. Lavellée and L.-P. Rivest, eds.) 377–382. Dunod, Paris.

BAFFETTA, F., CORONA, P. and FATTORINI, L. (2010). Design-based diagnostics for k-NN estimators of forest resources. *Can. J. For. Res.* **41** 59–72.

BAFFETTA, F., FATTORINI, L., FRANCESCHI, S. and CORONA, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sens. Environ.* **113** 463–475.

BARDSLEY, P. and CHAMBERS, R. L. (1984). Multipurpose estimation from unbalanced samples. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **33** 290–299.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.

BEAUMONT, J. F. and BOCCI, C. (2008). Another look at ridge calibration. *Metron* **66** 5–20.

BEAUMONT, J.-F., HAZIZA, D. and RUIZ-GAZEN, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika* **100** 555–569.

BICKEL, P. J. and FREEDMAN, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12** 470–482. MR0740906

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292. MR0731144

BREIDT, F. J., CLAESKENS, G. and OPSOMER, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* **92** 831–846. MR2234189

BREIDT, F. J. and OPSOMER, J. D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.* **28** 1026–1053. MR1810918

BREIDT, F. J. and OPSOMER, J. D. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Ann. Statist.* **36** 403–427. MR2387977

BREIDT, F. J., OPSOMER, J. D. and SANCHEZ-BORREGO, I. (2016). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *J. Amer. Statist. Assoc.* **111** 822–833. MR3538708

BREIDT, F. J., OPSOMER, J. D., JOHNSON, A. A. and RANALLI, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Surv. Methodol.* **33** 35–44.

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CARDOT, H., GOGA, C. and LARDIN, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electron. J. Stat.* **7** 562–596. MR3035266

CARDOT, H. and JOSSERAND, E. (2011). Horvitz–Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika* **98** 107–118. MR2804213

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. MR0445666

CHAMBERS, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Off. Stat.* **12** 3–32.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

DAHLKE, M., BREIDT, F. J., OPSOMER, J. D. and VAN KEILEGOM, I. (2013). Nonparametric endogenous post-stratification estimation. *Statist. Sinica* **23** 189–211. MR3076164

DATTA, G. S. and GHOSH, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Ann. Statist.* **19** 1748–1770. MR1135147

DEVILLE, J.-C. and GOGA, C. (2004). Estimation par régression par polynômes locaux dans des enquêtes sur plusieurs échantillons. In *Echantillonnage et Méthodes D'Enquêtes* (P. Ardilly, ed.) 156–162. Dunod, Paris.

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382.

ELLIOTT, M. R. and LITTLE, R. J. A. (2000). Model-based alternatives to trimming survey weights. *J. Off. Stat.* **16** 191–209.

FAY, R. E. and HERRIOT, R. A. (1979). Estimation of income from small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141. MR1091842

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

FULLER, W. A. (2002). Regression estimation for survey samples (with discussion). *Surv. Methodol.* **28** 5–23.

GHOSH, M. and RAO, J. N. K. (1994). Small area estimation: An appraisal. *Statist. Sci.* **9** 55–93. MR1278679

GOGA, C. (2004). Estimation de l'évolution d'un total en présence d'information auxiliaire: Une approche par splines de régression. *C. R. Math. Acad. Sci. Paris* **339** 441–444.

GOGA, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: Une approche non paramétrique par splines de régression. *Canad. J. Statist.* **33** 163–180. MR2193026

GUGGEMOS, F. and TILLÉ, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *J. Statist. Plann. Inference* **140** 3199–3212. MR2659847

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 361–374. MR0125612

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference and Prediction*. Springer, New York.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685.

ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96. MR0648029

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* **9** 1010–1019. MR0628756

LAZZERONI, L. C. and LITTLE, R. J. A. (1998). Random-effects models for smoothing poststratification weights. *J. Off. Stat.* **14** 61–78.

LI, X. and OPSOMER, J. D. (2006). Model averaging in survey estimation. In *Proceedings of the Section on Survey Research Methods*. Amer. Statist. Assoc., Alexandria, VA.

MCCONVILLE, K. (2011). Improved Estimation for Complex Surveys Using Modern Regression Techniques. Ph.D. thesis, Colorado State University.

MCCONVILLE, K. S. and BREIDT, F. J. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *J. Nonparametr. Stat.* **25** 745–763. MR3174295

MCROBERTS, R. E., NÆSSET, E. and GOBAKKEN, T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sens. Environ.* **128** 268–275.

MCROBERTS, R. E., TOMPPO, E. O. and NÆSSET, E. (2010). Advances and emerging issues in national forest inventories. *Scand. J. For. Res.* **25** 368–381.

MONTANARI, G. E. and RANALLI, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *J. Amer. Statist. Assoc.* **100** 1429–1442.

MONTANARI, G. E. and RANALLI, M. G. (2009). Multiple and ridge model calibration. In *Proceedings of Workshop on Calibration and Estimation in Surveys*. Statistics Canada, Ottawa, ON.

NÆSSET, E., BOLLANDSÅS, O. M., GOBAKKEN, T., GREGOIRE, T. G. and STÅHL, G. (2013). Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne lidar: A case study with post-stratification to provide "activity data". *Remote Sens. Environ.* **128** 299–314.

OPSOMER, J. D., BREIDT, F. J., MOISEN, G. G. and KAUERMANN, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *J. Amer. Statist. Assoc.* **102** 400–409. MR2370838

OPSOMER, J. D., CLAESKENS, G., RANALLI, M. G., KAUERMANN, G. and BREIDT, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 265–286. MR2412642

PARK, M. and FULLER, W. A. (2005). Towards nonnegative regression weights for survey samples. *Surv. Methodol.* **31** 85–93.

PARK, M. and FULLER, W. A. (2009). The mixed model for survey regression estimation. *J. Statist. Plann. Inference* **139** 1320–1331.

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley-Interscience, New York.

RAO, J. N. K. and SINGH, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling (Pkg: P57-85). In *ASA Proceedings of the Section on Survey Research Methods* 57–65. Amer. Statist. Assoc., Alexandria, VA.

ROBINSON, P. and SÄRNDAL, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya, Ser. B* **45** 240–248. MR0748468

RUEDA, M., SÁNCHEZ-BORREGO, I. and ARCOS, A. (2009). Mean estimation in the presence of change points. *Appl. Math. Lett.* **22** 1257–1261. MR2532550

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

SÁNCHEZ-BORREGO, I., RUEDA, M. and MUÑOZ, J. (2012). Nonparametric methods in sample surveys. Application to the estimation of cancer prevalence. *Qual. Quant.* **46** 405–414.

SÄRNDAL, C.-E. (2010). The calibration approach in survey theory and practice. *Surv. Methodol.* **33** 99–119.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SILVA, P. N. and SKINNER, C. J. (1997). Variable selection for regression estimation in finite populations. *Surv. Methodol.* **23** 23–32.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

TIPTON, J., OPSOMER, J. and MOISEN, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sens. Environ.* **139** 130–137.

TOTH, D. and ELTINGE, J. L. (2011). Building consistent regression trees from complex sample data. *J. Amer. Statist. Assoc.* **106** 1626–1636. MR2896862

WANG, L. (2009). Single-index model-assisted estimation in survey sampling. *J. Nonparametr. Stat.* **21** 487–504. MR2571724

WANG, J. C., OPSOMER, J. D. and WANG, H. (2014). Bagging non-differentiable estimators in complex surveys. *Surv. Methodol.* **40** 189–209.

WANG, L. and WANG, S. (2011). Nonparametric additive model-assisted estimation for survey data. *J. Multivariate Anal.* **102** 1126–1140. MR2805653

WU, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika* **90** 937–951. MR2024768

WU, C. F. J. and SITTER, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* **96** 185–193. MR1952731

ZHENG, H. and LITTLE, R. J. A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *J. Off. Stat.* **19** 99–117.

ZHENG, H. and LITTLE, R. J. A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Surv. Methodol.* **30** 209–218.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469