T E C H N I C A L R E P O R T

11004

Nonparametric endogenous post-stratification estimation

DAHLKE, M., JAY BREIDT, F., OPSOMER, J. and I. VAN KEILEGOM

I A P S T A T I S T I C S N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

NONPARAMETRIC ENDOGENOUS POST-STRATIFICATION ESTIMATION

Mark Dahlke[1], F. Jay Breidt[1], Jean D. Opsomer[1] and Ingrid Van Keilegom[2]

[1]*Colorado State University and* [2]*Université catholique de Louvain*

Abstract: Post-stratification is used to improve the precision of survey estimators when categorical auxiliary information is available from external sources. In natural resource surveys, such information may be obtained from remote sensing data classified into categories and displayed as maps. These maps may be based on classification models fitted to the sample data. Such "endogenous post-stratification" violates the standard assumptions that observations are classified without error into post-strata, and post-stratum population counts are known. Properties of the endogenous post-stratification estimator (EPSE) are derived for the case of sample-fitted nonparametric models, with particular emphasis on monotone regression models. Asymptotic properties of the nonparametric EPSE are investigated under a superpopulation model framework. Simulation experiments illustrate the practical effects of first fitting a nonparametric model to survey data before post-stratifying.

Key words and phrases: Monotone regression, smoothing, survey estimation.

1. Introduction

Post-stratification is a common method for improving the precision of survey estimators when categorical auxiliary information is available from sources external to the survey. In surveys of natural resources, auxiliary information may be obtained from remote sensing data, classified into categories and displayed as pixel-based maps. These maps may be constructed based on classification models fitted to sample data. Methods used by the US Forest Service in its Forest Inventory and Analysis program include post-stratification (PS) of the sample data based on categories derived from the sample data. Such "endogenous post-stratification" violates the standard post-stratification assumptions that observations are classified without error into post-strata, and post-stratum population counts are known. Breidt and Opsomer (2008) derived properties of the endogenous post-stratification estimator for the case of a sample-fitted generalized linear model, from which the post-strata are constructed by dividing the range of the model predictions into predetermined intervals. Design consistency of the endogenous post-stratification estimator was established under general unequal-probability sampling designs. Under a superpopulation model, consistency and asymptotic normality of the endogenous post-stratification estimator (EPSE) were established, showing that EPSE has the same asymptotic variance as the traditional post-stratified estimator with fixed strata. Simulation experiments demonstrated that the practical effect of first fitting a model to the survey data before post-stratifying is small, even for relatively small sample sizes.

The motivation for studying endogenous post-stratification came from methods used by the U.S. Forest Service in producing estimators for the Forest Inventory and Analysis (FIA; see Frayer and Furnival 1999). These methods rely on post-stratification using classification maps derived from satellite imagery and other ancillary information. Because the FIA data represent a source of high quality ground-level information of forest characteristics, there is a clear desire for being "allowed" to use them in calibrating, i.e. estimating, the classification maps, and hence to apply EPSE in FIA. The results in Breidt and Opsomer (2008) were considered to provide some "weak justification" for doing so (see Czaplewski 2010), but the fact that they are restricted to parametric models limits their applicability in the FIA context, where the methods being used are often nonparametric in nature (e.g. Moisen and Frescino 2002).

In this paper, we extend the EPSE methodology to the nonparametric estimation context. We show here that the superpopulation results obtained for EPSE by Breidt and Opsomer (2008) continue to hold in this case. We focus on the case where the underlying model is nonparametric but monotone, which is the most reasonable scenario in surveys since the model is used to divide the sample into homogeneous classes. Our theoretical results are valid for a general class of nonparametric estimators that includes kernel regression and penalized spline regression.

In the following section we give the definitions of the estimators we propose in this paper. The asymptotic results are given in Section 3. Section 4 examines some of the models and estimators satisfying the outlined conditions, and in Section 5 we present the results of a small simulation study. The proofs of the asymptotic results are collected in the Appendix.

2. Definition of the estimator

Consider a finite population $U_N = \{1, \ldots, i, \ldots, N\}$. For each $i \in U_N$, an auxiliary vector $\boldsymbol{x}_i$ is observed. A probability sample s of size n is drawn from U_N according to a sampling design $p_N(\cdot)$, where $p_N(s)$ is the probability of drawing the sample s. Assume $\pi_{iN} = \Pr\{i \in s\} = \sum_{s:i\in s} p_N(s) > 0$ for all $i \in U_N$, and define $\pi_{ijN} = \Pr\{i, j \in s\} = \sum_{s:i,j\in s} p_N(s)$ for all $i, j \in U_N$. For compactness of notation we will suppress the subscript N and write π_i, π_{ij} in what follows. Various study variables, generically denoted y_i, are observed for $i \in s$.

The targets of estimation are the finite population means of the survey variables, $\bar{y}_N = N^{-1} \sum_{U_N} y_i$. A purely design-based estimator (with all randomness coming exclusively from the selection of s) is provided by the Horvitz-Thompson estimator (HTE)

$$\bar{y}_\pi = \frac{1}{N} \sum_{i\in s} \frac{y_i}{\pi_i}.$$

Post-stratification (PS) and endogenous post-stratification are methods that take advantage of auxiliary information available for the population to improve the efficiency of design-based estimators. Following Breidt and Opsomer (2008), we first introduce some non-standard notation for PS that will be useful in our later discussion of endogenous PS. Using the $\{\boldsymbol{x}_i\}_{i\in U_N}$ and a real-valued function $m(\cdot)$, a scalar index $\{m(\boldsymbol{x}_i)\}_{i\in U_N}$ is constructed and used to partition U_N into H strata according to predetermined stratum boundaries $-\infty \leq \tau_0 < \tau_1 < \cdots < \tau_{H-1} < \tau_H \leq \infty$. Typically, $m(\cdot)$ will be the true relationship between a specific study variable z_i and the auxiliary variable/vector $\boldsymbol{x}_i$. We assume the following additive error model,

$$z_i = m(\boldsymbol{x}_i) + \sigma(\boldsymbol{x}_i)\epsilon_i, \tag{2.1}$$

where $\sigma^2(\boldsymbol{x}_i)$ is the unknown variance function, and $\mathrm{E}(\epsilon_i|\boldsymbol{x}_i) = 0, \mathrm{Var}(\epsilon_i|\boldsymbol{x}_i) = 1$. Breidt and Opsomer (2008) considered the particular case in which the index function $m(\cdot)$ is parameterized by a vector, $\boldsymbol{\lambda}$. We will write $m_\lambda(\boldsymbol{x}_i)$ in that case.

For exponents $\ell = 0, 1, 2$ and stratum indices $h = 1, \ldots, H$, define

$$A_{Nh\ell}(m) = \frac{1}{N} \sum_{i \in U_N} y_i^\ell \mathrm{I}_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}}$$

and

$$A^*_{Nh\ell}(m) = \frac{1}{N} \sum_{i \in U_N} y_i^\ell \frac{\mathrm{I}_{\{i \in s\}}}{\pi_i} \mathrm{I}_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}} \tag{2.2}$$

where $\mathrm{I}_{\{C\}} = 1$ if the event C occurs, and zero otherwise. In this notation, stratum h has population stratum proportion $A_{Nh0}(m)$, design-weighted sample post-stratum proportion $A^*_{Nh0}(m)$, and design-weighted sample post-stratum y-mean $A^*_{Nh1}(m)/A^*_{Nh0}(m)$. The traditional design-weighted PS estimator (PSE) for the population mean $\bar{y}_N = N^{-1} \sum_{i \in U_N} y_i$ is then

$$\begin{aligned} \hat{\mu}^*_y(m) &= \sum_{h=1}^{H} A_{Nh0}(m) \frac{A^*_{Nh1}(m)}{A^*_{Nh0}(m)} \\ &= \sum_{i \in s} \left\{ \sum_{h=1}^{H} A_{Nh0}(m) \frac{N^{-1} \pi_i^{-1} \mathrm{I}_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}}}{A^*_{Nh0}(m)} \right\} y_i = \sum_{i \in s} w^*_{is}(m) y_i, \quad (2.3) \end{aligned}$$

where the sample-dependent weights $\{w^*_{is}(m)\}_{i \in s}$ do not depend on $\{y_i\}$, and so can be used for any study variable.

For the important special case of equal-probability designs, in which $\pi_i = nN^{-1}$, we write

$$A_{nh\ell}(m) = \frac{1}{n} \sum_{i \in s} y_i^\ell \mathrm{I}_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}}.$$

In this case, the equal-probability PSE for the population mean $\bar{y}_N$ is

$$\hat{\mu}_y(m) = \sum_{h=1}^{H} A_{Nh0}(m) \frac{A_{nh1}(m)}{A_{nh0}(m)} = \sum_{i \in s} w_{is}(m) y_i,$$

where the weights $\{w_{is}(m)\}_{i \in s}$ are obtained by substituting nN^{-1} for π_i in (2.3).

In parametric PS, the vector $\boldsymbol{\lambda}$ is known. In parametric endogenous PS, the vector $\boldsymbol{\lambda}$ is not known and needs to be estimated from the sample $\{\boldsymbol{x}_i, z_i : i \in s\}$ using, for example, maximum likelihood estimation or estimating equations. Thus, $m_\lambda(\boldsymbol{x}_i)$ is estimated by $m_{\hat{\lambda}}(\boldsymbol{x}_i)$, and the endogenous post-stratification estimator (EPSE) for the population mean $\bar{y}_N$ is then defined as

$$\hat{\mu}_y^*(m_{\hat{\lambda}}) = \sum_{h=1}^{H} A_{Nh0}(m_{\hat{\lambda}}) \frac{A_{Nh1}^*(m_{\hat{\lambda}})}{A_{Nh0}^*(m_{\hat{\lambda}})} = \sum_{i\in s} w_{is}^*(m_{\hat{\lambda}}) y_i.$$

This parametric EPSE was studied in Breidt and Opsomer (2008). We consider now the case where $m(\cdot)$ is not assumed to follow a specific parametric shape. Again, m is typically the true regression relationship between a specific study variable z_i and an auxiliary variable/vector $\boldsymbol{x}_i$ as in model (2.1).

The estimator $\hat{\mu}_y^*(m)$ is infeasible, because $m(\cdot)$ is unknown. We can estimate $m(\cdot)$ from the sample $\{(\boldsymbol{x}_i, z_i) : i \in s\}$ by nonparametric regression, and in this article we will explicitly consider both kernel and spline-based methods. However, results should also apply to other nonparametric and semi-parametric fitting methods such as regression trees, neural nets, GAMs, etc. Writing $\hat{m}$ for the nonparametric estimator, the nonparametric endogenous post-stratified estimator is then defined as

$$\hat{\mu}_y^*(\hat{m}) = \sum_{h=1}^{H} A_{Nh0}(\hat{m}) \frac{A_{Nh1}^*(\hat{m})}{A_{Nh0}^*(\hat{m})}. \tag{2.4}$$

For the important special case of equal-probability designs, in which $\pi_i = nN^{-1}$, the equal-probability NEPSE for the population mean $\bar{y}_N$ is

$$\hat{\mu}_y(\hat{m}) = \sum_{h=1}^{H} A_{Nh0}(\hat{m}) \frac{A_{nh1}(\hat{m})}{A_{nh0}(\hat{m})} = \sum_{i\in s} w_{is}(\hat{m}) y_i.$$

In order to study the properties of this estimator, it is sufficient to consider the following simpler estimators

$$A_{\tau\ell}(\hat{m}) = \frac{1}{N} \sum_{i\in U_N} y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}}$$

and

$$A_{\tau\ell}^*(\hat{m}) = \frac{1}{N} \sum_{i\in U_N} \frac{\mathrm{I}_{\{i\in s\}}}{\pi_i} y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}},$$

for a generic boundary value $\tau \in \{\tau_0, \tau_1, \cdots, \tau_H\}$. For equal probability designs we write

$$A_{n\tau\ell}(\hat{m}) = \frac{1}{n}\sum_{i\in s} y_i^{\ell} \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}}.$$

The form of these estimators suggests the use of tools from empirical process theory, which we turn to in the next section.

3. Main results

3.1 Superpopulation model assumptions

Before we explicitly state the model assumptions for studying the NEPSE estimator, we need to introduce the concept of *bracketing number* of empirical process theory (van der Vaart and Wellner 1996). For any $\varepsilon > 0$, any class $\mathcal{G}$ of measurable functions, and any norm $\|\cdot\|_{\mathcal{G}}$ defined on $\mathcal{G}$, $N_{[\,]}(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is the bracketing number, i.e. the minimal positive integer M for which there exist ε-brackets $\{[l_j, u_j] : \|l_j - u_j\|_{\mathcal{G}} \leq \varepsilon,\ \|l_j\|_{\mathcal{G}}, \|u_j\|_{\mathcal{G}} < \infty,\ j = 1, \ldots, M\}$ to cover $\mathcal{G}$ (i.e. for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, M\}$ such that $l_j \leq g \leq u_j$).

We make the following superpopulation model assumptions. Assumption 3.1.1 gives conditions on the multivariate distribution of covariates $\{\boldsymbol{x}_i\}$, 3.1.2 assumes equal probability sampling, assumptions 3.1.3 and 3.1.4 specify conditions on the sample fit $\hat{m}(\cdot)$, and assumption 3.1.5 gives moment conditions.

Assumption 3.1.1. *The covariates $\{\boldsymbol{x}_i\}$ are independent and identically distributed random p-vectors with nondegenerate continuous joint probability density function $f(\boldsymbol{x})$ and compact support. The function $u \to \Pr(m(\boldsymbol{x}) \leq u)$ is Lipschitz continuous of order $0 < \gamma \leq 1$, and*

$$\Pr(m(\boldsymbol{x}) \leq \tau_{h-1}) < \Pr(m(\boldsymbol{x}) \leq \tau_h)$$

for $h = 1, \ldots, H$.

Assumption 3.1.2. *The sample s is selected according to an equal-probability design of fixed size n, with $\pi_i = nN^{-1} \to \pi \in [0, 1]$, as $N \to \infty$.*

Assumption 3.1.3. *The nonparametric estimator $\hat{m}(\cdot)$ satisfies*

$$\sup_{\boldsymbol{x}} |\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| = o(1), a.s.$$

Assumption 3.1.4. *There exists a space* $\mathcal{D}$ *of measurable functions that satisfies* $m \in \mathcal{D}$, $\Pr(\hat{m} \in \mathcal{D}) \to 1$, *as* $n \to \infty$, *and*

$$\int_0^\infty \sqrt{\log N_{[]}(\lambda, \mathcal{F}, \|\cdot\|_2)}\, d\lambda < \infty,$$

where $\mathcal{F} = \{\boldsymbol{x} \to I_{\{d(\boldsymbol{x}) \leq \tau\}} : d \in \mathcal{D}\}$.

Assumption 3.1.5. *Given* $[\boldsymbol{x}_i]_{i \in U_N}$, *the study variables* $[y_i]_{i \in U_N}$ *are conditionally independent of the post-stratification variables* $[z_i]_{i \in U_N}$, *and* $y_i \mid \boldsymbol{x}_i$ *are conditionally independent random variables with* $E(y_i^{2\ell} \mid \boldsymbol{x}_i) \leq K_1 < \infty$, *for* $\ell = 0, 1, 2$.

3.2 Central limit theorem

For $\ell = 0, 1, 2$, define $\alpha_{\tau\ell}(m) = \mathrm{E}(y_i^\ell \mathrm{I}_{\{m(\boldsymbol{x}_i) \leq \tau\}})$. We start this section with a crucial lemma, which shows that $A_{\tau\ell}(\hat{m})$ (which is difficult to handle since it contains the nonparametric estimator $\hat{m}(\boldsymbol{x}_i)$ inside an indicator function) is asymptotically equivalent to $\mathrm{E}(y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i) \leq \tau\}} \mid \hat{m}) + A_{\tau\ell}(m) - \alpha_{\tau\ell}(m)$.

Lemma 1. *Under Assumptions 3.1.1–3.1.5,*

$$A_{\tau\ell}(\hat{m}) - E(y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i) \leq \tau\}} \mid \hat{m}) - A_{\tau\ell}(m) + \alpha_{\tau\ell}(m) = o_p(N^{-1/2}) \tag{3.1}$$

and

$$A_{n\tau\ell}(\hat{m}) - E(y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i) \leq \tau\}} \mid \hat{m}) - A_{n\tau\ell}(m) + \alpha_{\tau\ell}(m) = o_p(n^{-1/2}) \tag{3.2}$$

for $\ell = 0, 1, 2$.

We are now ready to state the main result of the paper.

Theorem 1. *Under Assumptions 3.1.1–3.1.5,*

$$\left\{\frac{1}{n}\left(1 - \frac{n}{N}\right)\right\}^{-1/2} (\hat{\mu}_y(\hat{m}) - \bar{y}_N) \xrightarrow{d} N(0, V_{ym}),$$

where

$$V_{ym} = \sum_{h=1}^H \Pr\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\} \mathrm{Var}(y_i | \tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h).$$

The proofs of both results are deferred to the Appendix.

3.3 Variance estimation

For the estimation of the variance V_{ym} we follow Breidt and Opsomer (2008).

Theorem 2. *Define*

$$\hat{V}_{y\hat{m}} = \sum_{h=1}^{H} \frac{A_{Nh0}^2(\hat{m})}{A_{nh0}(\hat{m})} \frac{A_{nh2}(\hat{m}) - A_{nh1}^2(\hat{m})/A_{nh0}(\hat{m})}{A_{nh0}(\hat{m}) - n^{-1}}. \tag{3.3}$$

Under Assumptions 3.1.1–3.1.5,

$$\left\{ \frac{1}{n} \left(1 - \frac{n}{N} \right) \right\}^{-1/2} \hat{V}_{y\hat{m}}^{-1/2} (\hat{\mu}_y(\hat{m}) - \bar{y}_N) \xrightarrow{d} N(0,1).$$

The proof can be found in the Appendix.

4. Applying the results

The results in the previous sections are expressed under quite general conditions on the class $\mathcal{D}$ and on the estimator $\hat{m}$. We now give some particular models for the regression function m and some particular estimators $\hat{m}$ for which the conditions are satisfied. The underlying models we consider are at least partly monotone, which is reasonable in this context because the function m is used to split the data into homogeneous cells.

4.1 Monotone regression

Let

$$\mathcal{D} = \{ d : R_X \to I\!R : d \text{ monotone and } \sup_{x \in R_X} |d(x)| \leq K \}$$

for some $K < \infty$, where R_X is a compact subset of $I\!R$. Suppose for simplicity that the functions in $\mathcal{D}$ are monotone decreasing. Then, the class $\mathcal{F}$ defined in assumption 3.1.4 is itself a set of one-dimensional bounded and monotone functions, and hence we have that

$$\log N_{[\,]}(\lambda, \mathcal{F}, \|\cdot\|_2) \leq K_1 \lambda^{-1}$$

for some $K_1 < \infty$, by Theorem 2.7.5 in van der Vaart and Wellner (1996). It follows that the integral in assumption 3.1.4 is finite.

Let $\hat{m}$ be any estimator of m for which $\sup_{x\in R_X} |\hat{m}(x) - m(x)| = o(1)$ a.s. Then, provided the true regression function m is monotone and bounded, we have that $\Pr(\hat{m} \in \mathcal{D}) \to 1$ as $n \to \infty$. The estimator $\hat{m}$ does not need to be monotone itself, a classical local polynomial or spline estimator does the job. Hence, Theorem 1 applies in this case. Moreover, the case of generalized monotone regression functions, obtained by using e.g. a logit transformation works as well. See Subsection 4.4 for more details.

4.2 Partially linear monotone regression

Consider now

$$\begin{aligned}\mathcal{D} \;=\; & \{R_X \to I\!R : (\boldsymbol{x}_1^T, x_2)^T \to \beta^T \boldsymbol{x}_1 + d(x_2) : \beta \in B \subset I\!R^k \text{ compact}, \\ & d \text{ monotone}, \sup_{x_2 \in R_{X_2}} |d(x_2)| \le K\},\end{aligned}$$

where $R_X = R_{X_1} \times R_{X_2}$ is a compact subset of $I\!R^{k+1}$. Suppose for simplicity that all coordinates of an arbitrary $\boldsymbol{x}_1 \in R_{X_1}$ and $\beta \in B$ are positive. Divide B into $r = O(\lambda^{-2k})$ pairs (β_i^L, β_i^U) $(i = 1, \ldots, r)$ that cover the whole set B and are such that $\sum_{l=1}^k (\beta_{il}^U - \beta_{il}^L)^2 \le \lambda^4$. Similarly, divide R_{X_1} into $s = O(\lambda^{-2k})$ pairs $(\boldsymbol{x}_{1j}^L, \boldsymbol{x}_{1j}^U)$ $(j = 1, \ldots, s)$ that cover R_{X_1} and are such that $\sum_{l=1}^k (x_{1jl}^U - x_{1jl}^L)^2 \le \lambda^4$. Let $d_1^L \le d_1^U, \ldots, d_q^L \le d_q^U$ be the $q = O(\exp(K\lambda^{-1}))$ $\|\cdot\|_\infty$-brackets for the space of bounded and monotone functions (see Theorem 2.7.5 in van der Vaart and Wellner (1996)). Then, for each $\beta \in B$ and d monotone and bounded, there exist i, j and l such that for all $(\boldsymbol{x}_1, x_2) \in R_X$:

$$\begin{aligned}\ell_{ijl}^L(x_2) &:= I_{\{\beta_i^{UT} \boldsymbol{x}_{1j}^U + d_l^U(x_2) \le \tau\}} \\ &\le I_{\{\beta^T \boldsymbol{x}_1 + d(x_2) \le \tau\}} \\ &\le I_{\{\beta_i^{LT} \boldsymbol{x}_{1j}^L + d_l^L(x_2) \le \tau\}} := u_{ijl}^U(x_2).\end{aligned}$$

It is easy to see that the brackets $(\boldsymbol{x}_1, x_2) \to (\ell_{ijl}^L(x_2), u_{ijl}^U(x_2))$ are λ-brackets with respect to the $\|\cdot\|_2$-norm. The number of these brackets is bounded by $\lambda^{-4k} \exp(K\lambda^{-1})$, and hence the integral in assumption 3.1.4 is finite.

The estimator $\hat{m}$ can, as in the previous example, be chosen as any uniformly consistent estimator of m. Then, $\Pr(\hat{m} \in \mathcal{D}) \to 1$ provided the true regression function m belongs to $\mathcal{D}$. This shows that Theorem 1 also holds true for this case.

4.3 Single index monotone regression

Our next example concerns a single index model with a monotone link function. Let

$$\mathcal{D} = \{R_X \to I\!R : \boldsymbol{x} \to d(\beta^T \boldsymbol{x}) : \beta \in B \subset I\!R^k \text{ compact}, d \text{ monotone}, \sup_u |d(u)| \leq K\},$$

where R_X is a compact subset of $I\!R^k$. The treatment of this case is similar to that of the partial linear monotone regression model. We omit the details.

4.4 Generalized nonparametric monotone regression

The use of generalized linear models in EPSE was initially discussed in Breidt and Opsomer (2008), This approach enjoys the benefit of being able to handle categorical response variables, and has (in many cases) obvious and easily interpretable boundary values. Denote the conditional moments of z_i given x_i, where the covariate x_i is univariate for ease of presentation, by

$$\mathrm{E}(z_i|x_i) = \mu(x_i), \mathrm{Var}(z_i|x_i) = \sigma^2(x_i) := V(\mu(x_i)).$$

We consider the case when there exists a known monotone link function $g(\cdot)$, such that $g(\mu(x_i)) = m(x_i)$, following the framework of McCullagh and Nelder (1989). We can define the quasi-likelihood function $Q(\mu(x), z)$ which satisfies

$$\frac{\partial}{\partial \mu(x)} Q(\mu(x), z) = \frac{z - \mu(x)}{V(\mu(x))},$$

as in McCullagh and Nelder (1989). The function $m(x)$ can be estimated nonparametrically, as suggested by Green and Silverman (1994), and Fan, Heckman, and Wand (1995), among other authors.

We propose to approximate the function $m(x)$ locally by a pth-degree polynomial $m(x) \approx \beta_0 + \beta_1(x - x_i) + \cdots + \beta_p(x - x_i)^p$ and maximize the weighted quasi-likelihood to estimate the function $m(x)$ at each location x on the support of x_i as suggested by Fan, Heckman, and Wand (1995),

$$\sum_{i\in s}\frac{1}{\pi_i}Q(g^{-1}(\beta_0+\beta_1(x-x_i)+\cdots+\beta_p(x-x_i)^p),z_i)K_h(x_i-x), \tag{4.1}$$

where $K_h(\cdot)=\frac{1}{h}K(\cdot/h)$ and $K(\cdot)$ is a kernel function (for details, see Simonoff 1996, Silverman 1999).

We let $(\hat{\beta}_{0x},\hat{\beta}_{1x},\ldots,\hat{\beta}_{px})$ be the minimizer of (4.1). Then, the model fitted value of $m(x)$ is $\hat{m}(x)=\hat{\beta}_{0x}$, and $\hat{\mathrm{E}}(z|X=x)=g^{-1}(\hat{m}(x))=g^{-1}(\hat{\beta}_{0x})$. Again, we could retain the boundary values for variable z, $\{\tau_0,\tau_1,\ldots,\tau_H\}$, and define $A^*_{Nh\ell}(\hat{m})$ similar to (2.2),

$$A^*_{Nh\ell}(\hat{m})=\frac{1}{N}\sum_{i\in U_N}y_i^\ell\frac{\mathrm{I}_{\{i\in s\}}}{\pi_i}\mathrm{I}_{\{\tau_{h-1}<g^{-1}(\hat{m}(\boldsymbol{x}_i))\le\tau_h\}}, \tag{4.2}$$

for $l=0,1,2$. Given (4.2), a natural estimator for the population mean $\bar{y}_N$ is the same as (2.4). The verification of assumptions 3.1.3 and 3.1.4 is similar to the verification in Subsection 4.1 and is therefore omitted.

5. Simulations

The main goal of the simulation is to assess the design efficiency of the NEPSE relative to competing survey estimators. The simulations are performed in a setting that mimics a real survey, in which characteristics of multiple study variables are estimated using one set of weights. We consider several different sets of weights for estimation of a mean: the Horvitz-Thompson estimator (HTE) weights $\{n^{-1}\}_{i\in s}$, the PSE weights $\{w_{is}(m)\}_{i\in s}$, the NEPSE weights $\{w_{is}(\hat{m})\}_{i\in s}$, and the simple linear regression (REG) weights (e.g. Särndal et al. 1992, equation (6.5.12)). We use $H=4$ strata with fixed, known boundaries $\boldsymbol{\tau}=(-\infty,0.5,1.0,1.5,\infty)$ for PSE and NEPSE. The HTE does not use auxiliary information; the PSE uses auxiliary information with a known model; the REG uses auxiliary information with a fitted parametric model, and the NEPSE uses auxiliary information with a fitted nonparametric model. Specifically, we use a linear penalized spline with approximate degrees of freedom determined by the smoothing parameter (Ruppert et al. 2003, §3.13). For comparison, we obtained an additional set of weights by fitting a nonparametric model using the entire finite population. The results from this set of weights are very similar to the PSE and NEPSE results and are not included in the table.

We generate a population of size $N = 1000$ with eight survey variables of interest. The values $x_1, \ldots, x_N$ are independent and uniformly distributed on $(0, 1)$. The first variable, `ratio`, is generated according to a regression through the origin or ratio model (see e.g. Särndal et al. 1992, p.226), with mean $1 + 2(x - 0.5)$ and with independent normal errors with variance $2\sigma^2 x$. For the next six variables (y_i), we take their mean functions to be equal to

$$2\frac{g_k(x) - \min_{x\in[0,1]} g_k(x)}{\max_{x\in[0,1]} g_k(x) - \min_{x\in[0,1]} g_k(x)}$$

where

$$\begin{aligned}
\texttt{quad}: &\quad g_1(x) = 1 + 2(x - 0.5)^2 \\
\texttt{bump}: &\quad g_2(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2) \\
\texttt{jump}: &\quad g_3(x) = \{1 + 2(x - 0.5)\}I_{\{x\leq 0.65\}} + 0.65 I_{\{x>0.65\}} \\
\texttt{expo}: &\quad g_4(x) = \exp(-8x) \\
\texttt{cycle1}: &\quad g_5(x) = 2 + \sin(2\pi x) \\
\texttt{cycle4}: &\quad g_6(x) = 2 + \sin(8\pi x).
\end{aligned}$$

This means that the minimum is 0 and the maximum is 2 for each of the first seven mean functions. Finally, the eighth survey variable is

$$\texttt{noise:} \quad g_7(x) = 8.$$

Independent normal errors with mean zero and variance equal to σ^2 are then added to each of these mean functions. Note that the variance function for the `ratio` model is chosen so that, averaging over the covariate x, we have $\mathrm{E}[v(x)] = \sigma^2$. Thus, the heteroskedastic `ratio` variable and the remaining seven study variables all have the same variance, averaged over x.

For given values of σ, we fixed the population (that is, simulated N values for each of the eight variables of interest) and drew 1000 replicate samples of size n, each via simple random sampling without replacement from this fixed population. We constructed HTE and REG weights using standard methods.

We then computed the ratio of the MSE for each competing estimator to that of the NEPSE.

In the first simulation experiment, we consider in detail the case in which the PS variable follows a regression through the origin or ratio model (see e.g. Särndal et al. (1992), p. 226). We used the `ratio` variable as the PS variable and computed PSE weights with known $m(x) = 1 + 2(x - 0.5)$ and NEPSE weights with (approximately) 2 or 5 degrees of freedom (df) in the smoothing spline. The weights were then applied to the remaining seven study variables. We also varied the noise variance ($\sigma = 0.25$ or $\sigma = 0.5$). With 2 df, the smoothing spline yields the linear (parametric) fit, and thus corresponds to EPSE. Results for this case, presented in Table 1, are qualitatively similar to those in Table 1 of Breidt and Opsomer (2008) (the results are different because the earlier paper fits regression through the origin instead of simple linear regression, and uses different signal-to-noise ratios since the mean functions are not scaled to [0,2]).

Note that NEPSE dominates HTE in every case except `cycle4` (since NEPSE does not have enough df to capture the four cycles and so its estimate of the mean function is oversmoothed and nearly constant) and `noise`, where NEPSE fits an entirely superfluous model. REG beats NEPSE for `ratio`, where REG has the correct working model, and is slightly better for `bump`, which is highly linear over most of its range. REG is also slightly better for `cycle4` and for `noise`. NEPSE performs far better than REG for all of the other variables.

The effect of changing degrees of freedom in NEPSE is negligible in this example, since the true model for the PS variable is in fact linear. The effect of increasing noise variance is quite substantial, bringing the performance of all estimators closer together, as expected. Finally, NEPSE is essentially equivalent to the PSE in terms of design efficiency, even for $n = 50$, implying that the effect of basing the PS on a nonparametric regression instead of on stratum classifications and stratum counts known without error from a source external to the survey is negligible for moderate to large sample sizes.

In the second simulation, we fix $n = 100$, $df \approx 5$, $\sigma = 0.25$ and consider four different PS variables: `ratio`, `quad`, `bump`, and `cycle1`. Table 2 summarizes the design efficiency results as ratios of the MSE of the HTE, PSE(4), or REG over the MSE of the NEPSE(4). Overall, the behavior of the NEPSE is consistent

Response Variable	$df \approx$	($\sigma = 0.25$) NEPSE(4) versus HTE	PSE(4)	REG	($\sigma = 0.5$) NEPSE(4) versus HTE	PSE(4)	REG
`ratio`	2	4.98	1.01	0.74	2.19	1.02	0.91
	5	4.68	0.95	0.69	2.21	1.03	0.91
`quad`	2	2.34	1.03	2.56	1.62	1.05	1.75
	5	2.29	1.01	2.51	1.50	0.97	1.62
`bump`	2	3.22	1.00	0.94	1.88	1.00	0.95
	5	3.26	1.01	0.95	1.90	1.02	0.96
`jump`	2	2.19	1.00	1.80	1.40	0.99	1.26
	5	2.13	0.97	1.76	1.33	0.94	1.20
`expo`	2	1.88	0.99	1.17	1.29	1.01	1.07
	5	1.88	0.99	1.17	1.28	1.01	1.06
`cycle1`	2	3.10	1.04	1.56	1.97	1.03	1.26
	5	3.04	1.02	1.53	1.96	1.02	1.25
`cycle4`	2	0.96	1.00	0.92	0.98	1.02	0.95
	5	0.98	1.02	0.94	1.00	1.05	0.98
`noise`	2	0.93	1.00	0.96	0.92	1.00	0.96
	5	0.92	0.99	0.95	0.93	1.01	0.97

Table 1: Ratio of MSE of Horvitz-Thompson (HTE), post-stratification on 4 strata (PSE(4)), and linear regression (REG) estimators to MSE of nonparametric endogenous post-stratification estimator on 4 strata (NEPSE(4)). Numbers greater than one favor NEPSE. Based on `ratio` post-stratification variable in 1000 replications of simple random sampling of size $n = 50$ from a fixed population of size $N = 1000$. Replications in which at least one stratum had fewer than two samples are omitted from the summary: 4 reps at $df \approx 2$, $\sigma = 0.5$ and 33 reps at $df \approx 5$, $\sigma = 0.5$.

with expectations. NEPSE produces a large improvement in efficiency relative to the HTE for the variable on which the PS is based, and usually for other variables as well. NEPSE is as good or better (i.e. MSE ratio > 0.95) than REG in all but 12 of the 32 cases considered: NEPSE loses out in particular when the true model is linear or nearly so (`bump`). The `noise` variable shows that, when a variable is not related to the stratification variable, the efficiency is near that of the HTE (since the stratification is unnecessary).

We also assessed the coverage of approximate confidence intervals computed using the normal approximation from Theorem 1 and the variance estimator from Theorem 2. Coverage of nominal 95% confidence intervals, $\hat{\mu}_y(\hat{m}) \pm 1.96\{n^{-1}(1-nN^{-1})\hat{V}_{y\hat{m}}\}^{1/2}$, was consistently in the range of 93% to 96%.

PS Variable	Estimator	`ratio`	`quad`	`bump`	`jump`	`expo`	`cycle1`	`cycle4`	`noise`
	HTE	5.17	2.46	3.48	2.12	2.13	3.31	0.99	0.95
`ratio`	PSE(4)	0.98	1.03	1.02	0.97	1.01	1.02	1.00	1.00
	REG	0.71	2.49	0.97	1.70	1.19	1.64	0.90	0.97
	HTE	0.97	5.47	1.01	1.53	1.31	0.97	0.98	0.96
`quad`	PSE(4)	1.01	1.00	1.02	1.02	1.04	1.00	1.03	0.99
	REG	0.13	5.53	0.28	1.23	0.73	0.48	0.89	0.98
	HTE	4.07	1.93	4.13	2.02	2.30	2.70	1.13	0.95
`bump`	PSE(4)	1.27	1.33	0.76	1.07	1.11	0.96	1.05	1.00
	REG	0.56	1.95	1.15	1.62	1.29	1.34	1.03	0.97
	HTE	2.89	1.01	2.53	1.26	1.35	5.68	1.00	0.97
`cycle1`	PSE(4)	1.01	1.00	1.06	1.04	0.96	0.92	1.03	1.01
	REG	0.40	1.02	0.70	1.01	0.75	2.81	0.91	0.99

Table 2: Ratio of MSE of Horvitz-Thompson (HTE), post-stratification on 4 strata (PSE(4)), and linear regression (REG) estimators to MSE of nonparametric endogenous post-stratification estimator on 4 strata (NEPSE(4)). Numbers greater than one favor NEPSE. Based on four different PS variables in 1000 replications of simple random sampling of size $n = 100$ from a fixed population of size $N = 1000$.

Acknowledgment

Van Keilegom acknowledges financial support from IAP research network nr. P6/03 of the Belgian government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. The research of the remaining three authors was supported in part by the US National Science Foundation (SES-0922142).

Appendix

Proof of Lemma 1. The expression on the left hand side of (3.1) equals

$$N^{-1} \sum_{i \in U_N} \{y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i) \le \tau\}} - y_i^\ell I_{\{m(\boldsymbol{x}_i) \le \tau\}} - \mathrm{E}[y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i) \le \tau\}} \mid \hat{m}] + E[y_i^\ell I_{\{m(\boldsymbol{x}_i) \le \tau\}}]\}.$$

Let

$$\begin{aligned} \mathcal{H} &= \{(\boldsymbol{x}, y) \to y^\ell I_{\{d(\boldsymbol{x}) \le \tau\}} - y^\ell I_{\{m(\boldsymbol{x}) \le \tau\}} \\ &\quad - E[y^\ell I_{\{d(\boldsymbol{x}) \le \tau\}}] + E[y^\ell I_{\{m(\boldsymbol{x}) \le \tau\}}] : d \in \mathcal{D}\}, \end{aligned}$$

where $\mathcal{D}$ is defined as in assumption 3.1.4.

In a first step we will show that the class $\mathcal{H}$ is Donsker. From Theorem 2.5.6 in van der Vaart and Wellner (1996), it follows that it suffices to show that

$$\int_0^\infty \sqrt{\log N_{[]}(\lambda, \mathcal{H}, \|\cdot\|_2)}\, d\lambda < \infty. \tag{A.1}$$

From assumption 3.1.4 we know that the class

$$\mathcal{F} = \{(\boldsymbol{x}, y) \to y^\ell I_{\{d(\boldsymbol{x}) \leq \tau\}} : d \in \mathcal{D}\}$$

satisfies (A.1) with $\mathcal{H}$ replaced by $\mathcal{F}$, and hence the same holds for $\mathcal{H}$ itself, since the three other terms in $\mathcal{H}$ do not change its bracketing number.

Let

$$\hat{h}(\boldsymbol{x}, y) = y^\ell \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right) - \mathrm{E}\left[y^\ell \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right) \middle| \hat{m}\right],$$

where $(\boldsymbol{x}, y)$ is independent of the fit, $\hat{m}(\cdot)$. Then

$$\begin{aligned}
&\mathrm{Var}\left(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}\right) \\
&= \mathrm{Var}\left(y^\ell \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right) \middle| \hat{m}\right) \\
&\leq \mathrm{E}\left[\left(y^\ell \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right)\right)^2 \middle| \hat{m}\right] \\
&= \mathrm{E}\left[y^{2\ell} \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right)^2 \middle| \hat{m}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[y^{2\ell} \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right)^2 \middle| \hat{m}, \boldsymbol{x}\right] \middle| \hat{m}\right] \\
&= \mathrm{E}\left[\mathrm{E}[y^{2\ell} \mid \hat{m}, \boldsymbol{x}] \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right)^2 \middle| \hat{m}\right] \\
&= \mathrm{E}\left[\mathrm{E}[y^{2\ell} \mid \boldsymbol{x}] \left(I_{\{\hat{m}(\boldsymbol{x}) \leq \tau\}} - I_{\{m(\boldsymbol{x}) \leq \tau\}}\right)^2 \middle| \hat{m}\right] \\
&\leq K_1 \{\Pr(\hat{m}(\boldsymbol{x}) \leq \tau, m(\boldsymbol{x}) > \tau \mid \hat{m}) \\
&\quad + \Pr(\hat{m}(\boldsymbol{x}) > \tau, m(\boldsymbol{x}) \leq \tau \mid \hat{m})\},
\end{aligned} \tag{A.2}$$

where K_1 is given in assumption 3.1.5. Let $\epsilon > 0$ be given. By assumption 3.1.1, $F(u) = \Pr(m(\boldsymbol{x}) \leq u)$ is uniformly continuous, so there exists $\delta > 0$ such that $|u_1 - u_2| \leq \delta$ implies $|F(u_1) - F(u_2)| < \epsilon$. We will show that $\Pr(\hat{m}(\boldsymbol{x}) \leq \tau, m(\boldsymbol{x}) >$

$\tau \mid \hat{m}) = o_p(1)$. Consider

$$\begin{aligned}
&\Pr\Big(\Pr(\hat{m}(\boldsymbol{x}) \le \tau, m(x) > \tau \mid \hat{m}) > \epsilon\Big) \\
&\le \Pr\Big(\Pr(\hat{m}(\boldsymbol{x}) \le \tau, m(x) > \tau \mid \hat{m}) > \epsilon, \sup_{\boldsymbol{x}} |\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| \le \delta\Big) \\
&\quad + \Pr\Big(\sup_{\boldsymbol{x}} |\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| > \delta\Big) \\
&\le \Pr\Big(\Pr(m(\boldsymbol{x}) - \delta \le \tau, m(\boldsymbol{x}) > \tau \mid \hat{m}) > \epsilon\Big) + o(1) \\
&= \Pr\Big(\Pr(m(\boldsymbol{x}) - \delta \le \tau, m(\boldsymbol{x}) > \tau) > \epsilon\Big) + o(1) \\
&= I_{\{F(\tau+\delta) - F(\tau) > \epsilon\}} + o(1) = o(1), \qquad \text{(A.3)}
\end{aligned}$$

by choice of δ, where the second inequality follows from assumption 3.1.3. Similarly,

$$\Pr(\hat{m}(\boldsymbol{x}) > \tau, m(\boldsymbol{x}) \le \tau \mid \hat{m}) = o_p(1). \qquad \text{(A.4)}$$

For fixed $\eta > 0, \lambda > 0$ consider

$$\begin{aligned}
&\Pr\Big(N^{1/2}|A_{\tau\ell}(\hat{m}) - \mathrm{E}[y_i^\ell \mathrm{I}_{\{\hat{m}(\boldsymbol{x}_i) \le \tau\}} \mid \hat{m}] - A_{\tau\ell}(m) + \alpha_{\tau\ell}(m)| > \lambda\Big) \\
&= \Pr\left(N^{-1/2}\left|\sum_{i \in U_N} \hat{h}(\boldsymbol{x}_i, y_i)\right| > \lambda\right) \\
&\le \Pr\left(N^{-1/2}\left|\sum_{i \in U_N} \hat{h}(\boldsymbol{x}_i, y_i)\right| > \lambda, \mathrm{Var}(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}) < \eta, \hat{m} \in \mathcal{D}\right) \\
&\quad + \Pr\left(N^{-1/2}\left|\sum_{i \in U_N} \hat{h}(\boldsymbol{x}_i, y_i)\right| > \lambda, \mathrm{Var}(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}) \ge \eta, \hat{m} \in \mathcal{D}\right) \\
&\quad + \Pr(\hat{m} \notin \mathcal{D}) \\
&\le \Pr\left(\sup_{h \in \mathcal{H}, \mathrm{Var}(h) < \eta} N^{-1/2}\left|\sum_{i \in U_N} h(\boldsymbol{x}_i, y_i)\right| > \lambda\right) \\
&\quad + \Pr\Big(\mathrm{Var}(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}) \ge \eta\Big) + \Pr(\hat{m} \notin \mathcal{D}) \\
&= d_{1N} + d_{2N} + d_{3N}.
\end{aligned}$$

As $N \to \infty$, $d_{1N} = o(1)$ as $\eta \downarrow 0$ by Corollary 2.3.12 in van der Vaart and Wellner

(1996) and the fact that $\mathcal{H}$ is Donsker. Also, $d_{2N} = o(1)$ by the arguments in (A.2)–(A.4), and $d_{3N} = o(1)$ by assumption 3.1.4. This establishes (3.1), and similar arguments verify (3.2).

Proof of Theorem 1. Note that $A_{Nh\ell}(M) = A_{\tau_h\ell}(M) - A_{\tau_{h-1}\ell}(M)$ and $A_{nh\ell}(M) = A_{n\tau_h\ell}(M) - A_{n\tau_{h-1}\ell}(M)$, for $M = \{m, \hat{m}\}$. Let

$$\alpha_{h\ell}(m) = \alpha_{\tau_h\ell}(m) - \alpha_{\tau_{h-1}\ell}(m) = \mathrm{E}[y_i^\ell \mathrm{I}_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\leq\tau_h\}}].$$

Then, applying Lemma 1 to two consecutive boundary values, τ_{h-1} and τ_h, we have that the difference of the expressions is

$$A_{Nh\ell}(\hat{m}) - \mathrm{E}[y_i^\ell \mathrm{I}_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\leq\tau_h\}} \mid \hat{m}] - A_{Nh\ell}(m) + \alpha_{h\ell}(m) = o_p(N^{-1/2}), \quad \text{(A.5)}$$

and

$$A_{nh\ell}(\hat{m}) - \mathrm{E}[y_i^\ell \mathrm{I}_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\leq\tau_h\}} \mid \hat{m}] - A_{nh\ell}(m) + \alpha_{h\ell}(m) = o_p(n^{-1/2}). \quad \text{(A.6)}$$

Given (A.5) and (A.6), the remainder of the proof is very similar to the corresponding proof in Breidt and Opsomer (2008). We mention highlights of that proof (in the NEPSE context) and omit much of the detail. Begin by defining $a_h = A_{Nh0}(m) - A_{nh0}(m)$ and $b_h = A_{Nh1}(m) - A_{nh1}(m)$. Calculation of appropriate covariances shows that $a_h = O_p\left(n^{-1/2}\right)$ and $b_h = O_p\left(n^{-1/2}\right)$. By arguments similar to those in (A.2),

$$\begin{aligned}
&\mathrm{E}\left[\left\{\mathrm{E}[y_i^\ell \mathrm{I}_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\leq\tau_h\}} \mid \hat{m}] - \alpha_{h\ell}(m)\right\}^2\right] \\
&\leq \mathrm{E}\Big[K_1\Big\{\Pr(\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \leq \tau_h, m(\boldsymbol{x}_i) > \tau_h \mid \hat{m}) \\
&\quad + \Pr(\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \leq \tau_h, m(\boldsymbol{x}_i) \leq \tau_{h-1} \mid \hat{m}) \\
&\quad + \Pr(\hat{m}(\boldsymbol{x}_i) > \tau_h, \tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h \mid \hat{m}) \\
&\quad + \Pr(\hat{m}(\boldsymbol{x}_i) \leq \tau_{h-1}, \tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h \mid \hat{m})\Big\}\Big]. \qquad \text{(A.7)}
\end{aligned}$$

We want to show that (A.7) converges to 0 as $n \to \infty$. Note that for a given

$\epsilon > 0$,

$$\begin{aligned}&\Pr\Big(\Pr(\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \leq \tau_h, m(\boldsymbol{x}_i) > \tau_h \mid \hat{m}) > \epsilon\Big)\\ &\quad\leq \ \Pr\Big(\Pr(\hat{m}(\boldsymbol{x}_i) \leq \tau_h, m(\boldsymbol{x}_i) > \tau_h \mid \hat{m}) > \epsilon\Big) = o(1),\end{aligned}$$

by (A.3). Similar reasoning shows that each of the terms inside the expectation in (A.7) is $o_p(1)$. By uniform integrability, (A.7) is $o(1)$. Thus, $\mathrm{E}[y_i^{\ell}\mathrm{I}_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\leq\tau_h\}} \mid \hat{m}]$ converges to $\alpha_{h\ell}(m)$ in mean square, and hence in probability.

Next,

$$A_{Nh\ell}(m) - \alpha_{h\ell}(m) = O_p\left(N^{-1/2}\right) \text{ and } A_{nh\ell}(m) - \alpha_{h\ell}(m) = O_p\left(n^{-1/2}\right)$$

by the central limit theorem. Further note that $A_{nhl}(m)$ and $A_{Nhl}(m)$ are $O_p(1)$ by the weak law of large numbers.

Since $\alpha_{h0}(m) > 0$ by assumption 3.1.1, we have

$$\frac{1}{A_{nh0}(\hat{m})} = \frac{1}{\alpha_{h0}(m)} + o_p(1). \tag{A.8}$$

We substitute (A.5), (A.6), and (A.8), and apply the established order results to show that the NEPSE error,

$$\hat{\mu}_y(\hat{m}) - \bar{y}_N = \sum_{h=1}^{H}\left\{\frac{A_{Nh0}(\hat{m})A_{nh1}(\hat{m}) - A_{nh0}(\hat{m})A_{Nh1}(\hat{m})}{A_{nh0}(\hat{m})}\right\},$$

can be rewritten as

$$\begin{aligned}&\hat{\mu}_y(\hat{m}) - \bar{y}_N \qquad \text{(A.9)}\\ &= \sum_{h=1}^{H}\left\{\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\left(A_{Nh0}(m) - A_{nh0}(m)\right) - \left(A_{Nh1}(m) - A_{nh1}(m)\right)\right\} + o_p\left(n^{-1/2}\right),\end{aligned}$$

showing the asymptotic distribution is the same as that obtained when $m(\cdot)$ is known.

To derive the asymptotic distribution, we apply the central limit theorem to (A.9) and refer to previously mentioned covariance computations. The limiting distribution of the NEPSE error is normal with mean zero and the variance is approximated by

$$\begin{aligned}
&\operatorname{Var}(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \\
&\simeq -\frac{1}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + \frac{1}{n}\left(1-\frac{n}{N}\right)\left(\sum_{h=1}^{H}\alpha_{h1}(m)\right)^2 + \operatorname{Var}(\bar{y}_\pi - \bar{y}_N) \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + [\mathrm{E}(y_i)]^2 + \operatorname{Var}(y_i)\right\}.
\end{aligned}$$

By definition of expectation given an event,

$$\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} = \mathrm{E}[y_i \,|\, \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h]$$

and

$$\mathrm{E}(y_i^2) = \sum_{h=1}^{H}\alpha_{h0}(m)\left\{\operatorname{Var}(y_i \,|\, \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h) + [\mathrm{E}(y_i \,|\, \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h)]^2\right\},$$

from which the variance given in Theorem 1 immediately follows.

Proof of Theorem 2. With only notational changes to indicate NEPSE results, this proof is identical to the corresponding EPSE proof of Breidt and Opsomer (2008). Note that $A_{Nh\ell}(m) \xrightarrow{P} \alpha_{h\ell}(m)$ and $A_{nh\ell}(m) \xrightarrow{P} \alpha_{h\ell}(m)$ as $n, N \to \infty$ by the weak law of large numbers, and $\mathrm{E}[y_i^\ell \mathrm{I}_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \le \tau_h\}} \mid \hat{m}] \xrightarrow{P} \alpha_{h\ell}(m)$ for $\ell = 0, 1, 2$ by the arguments following (A.7). Using equations (A.5) and (A.6), the expression given for $\hat{V}_{y\hat{m}}$ in (3.3) converges in probability to

$$\sum_{h=1}^{H}\alpha_{h0}(m)\left\{\frac{\alpha_{h2}(m)}{\alpha_{h0}(m)} - \left(\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\right)^2\right\}$$

from which the result follows by Slutsky's Theorem and Theorem 1.

References

Breidt, F. J. and J. D. Opsomer (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics 36*, 403–427.

Czaplewski, R. L. (2010). Complex sample survey estimation in static state-space. Gen. Tech. Rep. RMRS-GTR-xxx (in press), U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*(429), 141–150.

Frayer, W. E. and G. M. Furnival (1999). Forest survey sampling designs: A history. *Journal of Forestry 97*, 4–8.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. Washington, D. C.: Chapman and Hall.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman and Hall.

Moisen, G. G. and T. S. Frescino (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling 157*, 209–225.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

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

Silverman, B. W. (1999). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall Ltd.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag Inc.

Department of Statistics, Colorado State University, Fort Collins, CO 80523, U.S.A.

E-mail: dahlkem7@gmail.com

Department of Statistics, Colorado State University, Fort Collins, CO 80523, U.S.A.

E-mail: jbreidt@stat.colostate.edu

Department of Statistics, Colorado State University, Fort Collins, CO 80523, U.S.A.

E-mail: jopsomer@stat.colostate.edu

Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.

E-mail: ingrid.vankeilegom@uclouvain.be