Journal of Survey Statistics and Methodology (2019) 7, 398–421

INFORMATION CRITERION FOR NONPARAMETRIC MODEL-ASSISTED SURVEY ESTIMATORS

ADDISON JAMES
LAN XUE*
VIRGINIA LESSER

Nonparametric model-assisted estimators have been proposed to improve estimates of finite population parameters. Flexible nonparametric models provide more reliable estimators when a parametric model is misspecified. In this article, we propose an information criterion to select appropriate auxiliary variables to use in an additive model-assisted method. We approximate the additive nonparametric components using polynomial splines and extend the Bayesian Information Criterion (BIC) for finite populations. By removing irrelevant auxiliary variables, our method reduces model complexity and decreases estimator variance. We establish that the proposed BIC is asymptotically consistent in selecting the important explanatory variables when the true model is additive without interactions, a result supported by our numerical study. Our proposed method is easier to implement and better justified theoretically than the existing method proposed in the literature.

KEYWORDS: Finite population; Model-assisted; Nonparametric; Splines; Survey; Variable selection.

1. INTRODUCTION

Nonparametric modeling has gained popularity in survey statistics because of its flexibility in modeling nonlinear relationships. These models are applied to model-assisted estimation, which uses auxiliary variables observed on the

Addison James was a graduate student in the Department of Statistics, Oregon State University, Corvallis, OR 97331, USA, at the time this research was performed. Lan Xue is associate professor and Virginia Lesser is professor in the Department of Statistics, Oregon State University. This work was supported by Simons foundation (272556). We are grateful to the editor, the associate editor, and two referees whose comments led to significant improvements in the quality of the article.
*Address correspondence to Lan Xue, Department of Statistics, Oregon State University, Corvallis, OR 97331, USA; E-mail: xuel@science.oregonstate.edu.

doi: 10.1093/jssam/smy015 Advance access publication 19 June 2018

entire population to augment estimates of finite population quantities. Such auxiliary information is often available in natural resources and business surveys where business register or economic census provides access to extensive auxiliary information such as sales and employment on the entire population. However, model-assisted estimation is often impractical for personal, household or social science surveys where auxiliary information is not observed for every element of the population of interest. Dorfman (1992) first incorporated nonparametric regression with model-assisted estimation to improve the finite population estimates. Breidt and Opsomer (2000) extended univariate local polynomial regression to model-assisted estimators and proved that the estimate of the total is asymptotically design-unbiased and consistent. Breidt, Claeskens, and Opsomer (2005) proposed a class of univariate estimators based on penalized polynomial splines using a data-driven penalty parameter.

However, there are challenges in estimating nonparametric functions with a large number of predictor variables due to the "curse of dimensionality." Stone (1985) and Hastie and Tibshirani (1986) partially alleviate this by assuming the contribution of each covariate to be additive, but the form of each contribution is an unspecified univariate function. Opsomer, Breidt, Moisen, and Kauermann (2007) applied this additive model to model-assisted estimation using forest survey data.

In practice, a large number of auxiliary variables are available in many surveys, such as environmental surveys augmented by satellite data or health surveys with access to population registers. Model complexity can be reduced and the precision of the total estimate improved by using variable selection tools to select only important variables.

Even with a moderate number of variables, the large number of candidate models is a challenge for variable selection. An alternative to an exhaustive search is stepwise deletion or subset selection using criteria such as Mallows's C_p (Mallows 1973), Akaike Information Criterion (AIC, Akaike 1974), or Bayesian Information Criterion (BIC, Schwarz 1978). These criteria have previously been applied to data sampled independently from an infinite population.

Assuming a linear model, forms of the AIC and BIC have been derived for data obtained from complex sampling designs. Hens, Aerts, and Molenberghs (2006) proposed a form of AIC when there are missing observations for a single-stage design. An approximation to the BIC for complex sample designs was proposed by Fabrizi and Lahiri (2007). Xu, Chen, and Mantel (2013) gave an alternative BIC for survey data with a non-Bayesian justification. Lumley and Scott (2015) derived versions of AIC and BIC for complex designs.

Outside the survey sampling framework, variable selection approaches have been adapted to nonparametric models for data sampled from infinite populations. Chen and Tsay (1993) extended the idea of best subset regression to additive models for selecting lagged variables in time series models. Huang and Yang (2004) generalized AIC and BIC to nonparametric models estimated

with splines. Xue (2009) introduced consistent variable selection for the additive model using penalized polynomial splines. Huang, Horowitz, and Wei (2010) applied the adaptive LASSO to additive models and proved that it is a consistent method of variable selection. Wang and Wang (2011) proposed a BIC for variable selection in additive models with design-based samples. However, the asymptotic theory of variable selection to samples using unequal selection probabilities has not been examined.

In this article, we extend the information criterion of Huang and Yang (2004) to samples from finite populations. We propose a BIC for consistent variable selection in additive model-assisted estimation. Our proposed method is applicable for data generated from a broad range of survey designs. However, it is challenging to establish the consistency of the proposed variable selection method under the design-based survey framework. One difficulty arises from the fact that the nonparametric model is approximated using a finite set of parameters that increase in number as a function of sample size. Another difficulty is the need to account for two sources of variation: the probability sampling design and the data generating process from the superpopulation model. The variable selection method proposed by Wang and Wang (2011), like ours, assumes an additive model and is applicable to data sampled from finite populations. However, it differs from our method in that ours is based on the likelihood function rather than the asymptotic mean squared error as in Wang and Wang (2011). Furthermore, our method is consistent for complex designs beyond simple random sampling (SRS). Our numerical studies suggest that our method is better than that of Wang and Wang (2011) for small sample sizes.

The article is organized as follows: In section 2, we introduce additive model-assisted estimation using polynomial splines. Section 3 presents a derivation of the BIC and its consistency theorem, which is proved in the appendix. Simulation results under two sampling schemes and four superpopulation models are discussed in section 4. In Section 5, we apply our method to the California Academic Performance Index data set. Conclusions are given in section 6.

2. ADDITIVE MODEL-ASSISTED ESTIMATION

Model-assisted estimation incorporates auxiliary information along with design weights at the estimation stage by considering the finite population as a realization from a superpopulation (Särndal, Swensson, and Wretman 1992). A linear model is typically assumed. In this article, we assume the superpopulation model has a nonparametric additive form to capture possible nonlinear relationships between the auxiliary information and variable of interest. In this section, we introduce the additive model for model-assisted estimation and show how to estimate the model using polynomial splines.

2.1 The Model

Let $U_N = \{1, \ldots, N\}$ be a finite population. A sample, S, of fixed-size n_N is drawn from U_N using a probability sampling design D_N. Assume auxiliary information $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})'$ is known for all $i \in U_N$. The variable of interest, y_i, is known only for the elements sampled from the population. Define $I_i = 1$ if $i \in S$ and 0 otherwise. We denote the first order inclusion probability as $\pi_{i,N} = P_{D_N}(i \in S) = P_{D_N}(I_i = 1)$ and the second order inclusion probability as $\pi_{ij,N} = P_{D_N}(i, j \in S) = P_{D_N}(I_i I_j = 1)$. In the following, the second order inclusion probability can be unknown, but it needs to satisfy assumption (A2) in section 3 to guarantee desirable statistical proprieties of the proposed method. The subscript N will be omitted to simplify the notation.

Our objective is to estimate the finite population total $t_y = \sum_{i \in U} y_i$. Model-assisted estimation uses auxiliary information at the estimation stage by considering $\{Y_i, \mathbf{X}_i\}_{i=1}^N$ as i.i.d. realizations from a superpopulation, ξ, written as

$$Y_i = m(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \ldots, N,$$

where m is the true relationship between the variable of interest and the auxiliary variables, and the errors $\{\epsilon_i\}_{i=1}^N$ are independent and identically distributed with mean zero. In addition, $\{\epsilon_i\}_{i=1}^N$ and $\{\mathbf{X}_i\}_{i=1}^N$ are independent. The model-assisted estimator takes the form,

$$\hat{t}_{MA} = \sum_{i \in U} \hat{m}(\mathbf{x}_i) + \sum_{i \in S} \frac{y_i - \hat{m}(\mathbf{x}_i)}{\pi_i}, \tag{1}$$

where $\hat{m}$ is an estimate of m using the available sample. The model-assisted estimator in (1) takes advantage of the known auxiliary information to produce more efficient estimates. Särndal et al. (1992) contains a comprehensive overview of model-assisted estimators.

As in Hastie and Tibshirani (1986), we assume an additive nonlinear relationship between the auxiliary information and the variable of interest. That is,

$$m(\mathbf{X}) = \alpha_0 + \sum_{l=1}^{d} \alpha_l(X_l), \tag{2}$$

where α_0 is an unknown constant and $\{\alpha_l\}_{l=1}^d$ are unknown smooth univariate functions. For identifiability purposes and without loss of generality, it will be assumed that $X_l \in [0, 1]$ and $E[\alpha_l(X_l)] = 0$, for $l = 1, \ldots, d$. Compared with a classic linear model, the additive model can improve the efficiency of the total estimator because its nonparametric components can be more robust to model misspecification.

A major challenge in estimating nonparametric functions with more than one variable is dealing with the "curse of dimensionality." However, the additive structure in (2) allows the estimation of the additive model with the same optimal rate of convergence as the univariate case (Stone 1985). This article will develop methods and results for estimation from complex surveys only for additive models without interaction terms.

2.2 Polynomial Splines

With an appropriate choice of knots, polynomial splines often provide accurate approximations of smooth functions and have better convergence rates than regular polynomials without knots (see De Boor 2001, p. 149). To define them, let $C^p([0,1])$ be the space of p-times continuously differentiable functions, for some integer $p>0$. For each auxiliary variable $l=1,\ldots,d$, define a knot sequence $\kappa_{ln}=\{0=k_{l0}<k_{l1}<\cdots<k_{lJ_n}<k_{lJ_n+1}=1\}$ where J_n is the number of interior knots. Denote $\phi_l=\phi^p([0,1],\kappa_{l,n})\subset C^{p-1}([0,1])$ as the space of polynomial splines that are piece-wise polynomials of degree p or less on the intervals $\left[k_{l(i-1)},k_{li}\right),i=1,\ldots,J_n$ and $[k_{lJ_n},k_{l(J_n+1)}]$, and connect smoothly at the knots such that they are $(p-1)$ times continuously differentiable on $[0,1]$.

For a fixed p and J_n, let

$$\Gamma_l^*(X_l)=(X_l,\ldots,X_l^p,(X_l-k_{l1})_+^p,\ldots,(X_l-k_{lJ_n})_+^p)',$$

with $(x)_+=x$ if $x>0$, else $(x)_+=0$. With the intercept term, this is the degree p truncated power basis for the spline space ϕ_l with J_n knots. For $1\le j\le J_n+p$, let $\Gamma_{lj}^*(X_l)$ be the jth element of the vector $\Gamma_l^*(X_l)$. Define the centered basis $\Gamma_l(X_l)=(\Gamma_{l1}(X_l),\ldots,\Gamma_{l(J_n+p)}(X_l))'$, where $\Gamma_{lj}(X_l)=\Gamma_{lj}^*(X_l)-\hat{N}^{-1}\sum_{i\in S}\pi_i^{-1}\Gamma_{lj}^*(X_{li})$ with $\hat{N}=\sum_{i\in S}\pi_i^{-1}$. The basis is centered by the survey weighted means to consistently estimate the additive function α_l in (2). The centered basis for all d variables $\mathbf{X}=(X_1,\ldots,X_d)$ is then,

$$\Gamma(\mathbf{X})=(1,\Gamma_1(X_1)',\ldots,\Gamma_d(X_d)')'.$$

Suppose each additive component can be approximated by

$$\alpha_l(X_l)\approx g_l(X_l)=\sum_{j=1}^{J_n+p}\theta_{lj}\Gamma_{lj}(X_l).$$

Define $g(\mathbf{x})=\theta_0+\sum_{l=1}^d g_l(x_l)$ as the spline approximation of m. Let $\boldsymbol{\theta}_l=(\theta_{l1},\ldots,\theta_{l(J_n+p)})'$ be the J_n+p parameter vector for g_l. The unknown coefficients

$$\boldsymbol{\theta}=(\theta_0,\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_d')'$$

can then be estimated simultaneously by least squares.

For known $\mathbf{y} = (y_1, \ldots y_N)'$, the population estimate of $\boldsymbol{\theta}$ is

$$\tilde{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(y_i - \theta_0 - \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \theta_{lj} \Gamma_{lj}(x_{li}) \right)^2,$$

and

$$\tilde{\boldsymbol{\theta}} = [\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}\Gamma y,$$

where $\boldsymbol{\Gamma} = \left(\Gamma(\mathbf{x}_1), \ldots, \Gamma(\mathbf{x}_N) \right)'$ is the design matrix for the truncated power basis using the entire population. The basis in $\boldsymbol{\Gamma}$ is centered by survey-weighted sample means, instead of population means, so that the same basis can be used in both population and sample estimates defined in (3) and (4), respectively. The spline basis is centered to ensure consistent estimation of the additive functions in (2), which are of zero means. However this centering step only affects the constant term ascribed to each additive function and does not affect the variable selection. Therefore the spline can be estimated without the centering step if one is only interested in selecting relevant auxiliary variables. For a fixed $\mathbf{x}$, the population based estimate of m is given as

$$\tilde{m}(\mathbf{x}) = \tilde{\theta}_0 + \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \tilde{\theta}_{lj} \Gamma_{lj}(x_l). \tag{3}$$

In practice, since only the sampled values of the variable of interest are observed, $\mathbf{y}_S = \{y_i, i \in S\}$, an appropriate sample estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} N^{-1} \sum_{i \in S} \pi_i^{-1} \left(y_i - \theta_0 - \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \theta_{lj} \Gamma_{lj}(x_{li}) \right)^2,$$

and

$$\hat{\boldsymbol{\theta}} = [\boldsymbol{\Gamma}_S' \boldsymbol{\Pi}_S^{-1} \boldsymbol{\Gamma}_S]^{-1} \boldsymbol{\Gamma}_S' \boldsymbol{\Pi}_S^{-1} \mathbf{y}_S,$$

where $\boldsymbol{\Pi}_S^{-1} = \text{diag}(\{\pi_i^{-1}\}_{i \in S})$ and $\boldsymbol{\Gamma}_S = (\Gamma(\mathbf{x}_i)', i \in S)'$ is the design matrix for the truncated power basis using only the sample data. For given $\mathbf{x}$, the resulting sample estimate of m is

$$\hat{m}(\mathbf{x}) = \hat{\theta}_0 + \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \hat{\theta}_{lj} \Gamma_{lj}(x_l). \tag{4}$$

In practice, the estimate $\tilde{m}$ is not available since we do not observe every element of the population, but it serves as the population expected value of $\hat{m}$.

These expressions are useful for understanding the asymptotic properties of the estimator, as discussed in the appendix.

In addition, the estimated model $\hat{m}$ and the performance of the corresponding model-assisted estimator depend on the set of auxiliary variables included in the model and the knot sequence used in the spline approximation. In this article, we focus on the selection of auxiliary variables. For simplicity, we used the same sequence of knots for each covariate. The number of knots was chosen as the integer part of $n^{\frac{1}{2p+3}}$, the optimal rate of the knot number. More flexible approaches apply different knot sequences for each variable to provide different degrees of smoothness or can vary in both location and number of knots, depending on the data.

Wang and Wang (2011) suggested a similar method for estimating the total using the spline-backfitted local linear (SBLL) estimate of the additive model in (1). The SBLL estimator has two stages. The first stage applies polynomial spline regression to generate a pilot estimate, which is then used to construct pseudo-response values for each auxiliary variable. At the second stage, univariate local polynomial smoothing is applied to each pseudo-response and auxiliary variable pair. The resulting model-assisted estimator is asymptotically design unbiased, consistent, can be written as a weighted sum of calibrated weights (see Särndal et al. 1992), and asymptotically attains the Godambe-Joshi lower bound (Godambe and Joshi 1965). The authors proposed a "BIC-based method" of variable selection based on the asymptotic mean squared error (AMSE) and stated without proof that it is consistent under SRS.

The two-stage SBLL method has superior properties for estimating the additive components, but is computationally intensive since local polynomial smoothing is conducted on each variable in every model. Our goal focuses on variable selection, rather than estimation. We use only a single step of polynomial spline estimates and reduce the computation for each model.

3. PROPOSED INFORMATION CRITERION

Consider a set of candidate models $\{M_k\}$. For example, in an exhaustive search, the set includes all possible subsets of d candidate auxiliary variables. Suppose M_k contains d_k auxiliary variables, and the splines for $\{M_k\}$ involve a vector of parameters $\boldsymbol{\theta}_k$ with length $q_{k,n} = d_k(J_n + p)$, where J_n is the number of interior knots in the spline approximation. Let $L_S(\boldsymbol{\theta}_k)$ be the pseudo-log-likelihood function defined as $L_S(\boldsymbol{\theta}_k) = \frac{n}{N}\sum_{i\in S}\pi_i^{-1} l_i(\boldsymbol{\theta}_k)$, where $l_i(\boldsymbol{\theta}_k)$ is the contribution to the log-likelihood function from the i-th element in the sample. It is the design-weighted version of the census log-likelihood function, which is available only when every element in the population is observed and is defined as $L_N(\boldsymbol{\theta}_k) = \sum_{i\in U} l_i(\boldsymbol{\theta}_k)$. The scaling constant $\frac{n}{N}$ ensures that $n^{-1}L_S(\boldsymbol{\theta}_k)$ is design-unbiased to $N^{-1}L_N(\boldsymbol{\theta}_k)$. We propose the following nonparametric

BIC criterion for the selection of auxiliary variables in model-assisted estimation

$$\mathrm{BIC}(M_k) = -2L_S(\hat{\boldsymbol{\theta}}_k) + q_{k,n} \log n. \tag{5}$$

The pseudo-log-likelihood $L_S(\hat{\boldsymbol{\theta}}_k)$ in our proposed BIC incorporates sampling weights. A similar idea is also used in Xu et al. (2013) for linear models. For SRS, this formula reduces to the standard nonparametric BIC in Huang and Yang (2004), which is developed for data independently sampled from an infinite population. It is desirable because the simple random sample from the finite population can also be regarded as a random sample from the superpopulation.

Although we have focused on the pseudo-log-likelihood in (5), there are other approaches. For example, a similar BIC criterion can be constructed using the exact log-likelihood for the sample (Krieger and Pfeffermann 1992). Unlike the pseudo-likelihood, the calculation of the sample likelihood is often complicated, and it requires not only the sampling weights but also the full specification of the sampling mechanism. In addition, instead of the rescaled pseudo-log-likelihood in (5), one can also use the census log-likelihood estimated as $\sum_{i\in S} \pi_i^{-1} l_i(\boldsymbol{\theta}_k)$ and adjust the penalty term to be $q_{k,n} \log N$ in (5). The resulting BIC is asymptotically equivalent to (5) when n/N is bounded away from 0, and imposes a less stringent penalty on the model complexity when $n/N \to 0$. An in-depth comparison of the proposed BIC in (5) with the two possible BICs mentioned here is worth further investigation.

For normal model errors *with equal variances*, (5) becomes

$$\mathrm{BIC}(M_k) = \frac{n}{N}\left(\sum_{i\in S} \pi_i^{-1}\right) \log(\mathrm{WMSE}_{\hat{M}_k}) + q_{k,n} \log n, \tag{6}$$

where the weighted mean squared error (WMSE) for candidate model M_k is defined as

$$\mathrm{WMSE}_{\hat{M}_k} = \frac{\sum_{i\in S} \pi_i^{-1} (y_i - \hat{m}_{M_k}(\mathbf{x}_i))^2}{\sum_{i\in S} \pi_i^{-1}}. \tag{7}$$

Equation (6) is similar to the definition of the AIC in Hens et al. (2006).

To develop the theoretical properties of the proposed BIC, we introduce the following assumptions:

(A1) There exists a constant $B > 0$ such that $P\left(\limsup_{N\to\infty} \frac{1}{N}\sum_{i\in U} \left(\frac{\epsilon_i}{\pi_i}\right)^4 \le B\right) = 1$.

(A2) $\limsup_{N\to\infty} \max_{i\in U} \left\{\sum_{j\in U} (\pi_{ij} - \pi_i\pi_j)_-\right\} < \infty$, where $x_- = \max(0, -x)$.

(A3) For all $i \in U$, ε_i are independent and identically distributed. In addition, ε_i is independent of $\mathbf{X}_i$ with $E[\varepsilon_i] = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2, E[\varepsilon_i^4] = \mu_4 < \infty$.

(A4) For $i \in U$, the auxiliary variables $\mathbf{X}_i$ are independent and follow the same distribution as $\mathbf{X}$. Without loss of generality, we assume the support of $\mathbf{X}$ is $[0,1]^d$. Furthermore, the probability density function of X is bounded away from 0 and infinity on the support, written as $0 < f_X(\mathbf{x}) < \infty, \forall \mathbf{x} \in [0,1]^d$.

(A5) The number of knots is asymptotically related to the sample size such that $J_n \asymp n^{1/(2p+3)}$ and the spacing of the knots, $k_1, \ldots, k_{J_n}$, is such that $\min_{j \in \{1,\ldots,J_{n-1}\}} |k_{j+1} - k_j| / \max_{j \in \{1,\ldots,J_{n-1}\}} |k_{j+1} - k_j| > c$ for some constant $c > 0$.

(A6) Let M_0 be the indices of the auxiliary variables in the true model. We assume $\alpha_l \in \mathbb{C}^{p+1}[0,1]$ for $l \in M_0$, where $\mathbb{C}^{p+1}[0,1]$ denotes the space of $(p+1)$-times continuously differentiable functions.

Assumptions (A1) and (A2) ensure the consistency of the Horvitz-Thompson estimators under a fixed-size sampling design. When π_i is lower bounded with $\lambda_\pi \leq \liminf_{N \to \infty} \min_{i \in U} \pi_i$ for a constant $\lambda_\pi > 0$, then (A1) is satisfied under (A3). For SRS and many other types of sampling design, (A1) requires the sampling fraction $\frac{n}{N}$ to be bounded away from zero. Assumptions (A3) and (A4) make general assumptions about the superpopulation model errors and auxiliary variables. The most important feature of (A4) is assuming a compact support. Without loss of generality, data on any bounded interval can be rescaled to unit length. Assumption (A5) ensures the number of knots increase at an appropriate rate. Assumptions (A3)–(A6) are common in nonparametric estimation literature.

Theorem 1: Let M_0 be the indices of the auxiliary variables in the true model. Under assumptions (A1)–(A6), for any $1 > \epsilon > 0$ with superpopulation-model probability approaching 1 as both N and n approach infinity,

$$P_D\Big(\mathrm{BIC}(M_0) \leq \mathrm{BIC}(M), \text{ for all } M \neq M_0, M \subset \{1, \ldots, d\}\Big) \geq 1 - \epsilon,$$

where P_D denotes the design based probability measure.

Theorem one states that under regularity conditions, the true model has the lowest BIC among all candidate models with high design probability given the superpopulation when the population (and sample) size go to infinity. Searching the model space by calculating the BIC values for all models is often impractical due to the size of the space. A common approach is to use either forward or backward stepwise selection procedure instead of an exhaustive

search. Although there is no theoretical guarantee, these procedures have promising finite sample performance as shown in our simulation studies.

4. SIMULATION RESULTS

This section summarizes results from a simulation study that evaluates the performance of our proposed model selection criterion under different scenarios. We consider different regression models with various shapes of the additive components and different noise levels, and use polynomial splines with different degrees of smoothness. We are also interested in comparing the accuracy of variable selection of our method with the method proposed in Wang and Wang (2011) and investigating the effect of variable selection on the estimation of population totals. Our simulation results suggest the proposed method performs better than the method in Wang and Wang (2011) in selecting the correct set of auxiliary variables. The data is generated using SRS. An example using stratified sampling is included in the supplementary data online.

The setup of the simulation is identical to that used by Wang and Wang (2011). Observations are generated from four superpopulation models.

1. $Y = -1 + 2X_3 + 4X_6 + \sigma_0\varepsilon,$
2. $Y = 5.5 - 6X_2 + 8(X_2 - .5)^2 - 3X_{10} + 32(X_{10} - .5)^3 + \sigma_0\varepsilon,$
3. $Y = 8(X_2 - .5)^2 + \exp(2X_5 - 1) + 2\sin\{2\pi(X_8 - .5)\} + \sigma_0\varepsilon,$
4. $Y = \sum_{\alpha=1}^{5} \sin\{2\pi(X_\alpha - .5)\} + \frac{\sigma_0}{2}(\sum_{\alpha=1}^{5} X_\alpha)^{1/2}\varepsilon.$

In all four models, $\{\varepsilon_i\}_{i=1}^{N}$ are independently standard normal, and auxiliary variables $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,10})^{\mathrm{T}}$ are independently generated from the [Uniform(0, 1)]^{10} distribution for $i = 1, \ldots, N = 1000$. The scale parameter, σ_0, takes value 0.1 or 0.4. Simple random samples of size $n = 50$, 100, and 200 are drawn without replacement from the finite population. In this example, the auxiliary variables are generated independently of each other, which usually gives favorable variable selection results compared with the case when auxiliary variables are correlated.

For each sample, the proposed variable selection method was applied. The model with the lowest BIC score is called the *selected model*. The selected model is then used to estimate the finite population total of interest, t_y, using (1). For comparison, we also estimate t_y using the Horvitz-Thompson estimator $\hat{t}_{y,HT} = \sum_{i\in s} y_i/\pi_i$, where π_i is the first order inclusion probability of element i (see section 2.1).

The additive model is estimated using both linear and quadratic splines with knots spaced evenly between zero and one. The number of interior knots is two for linear splines and one for quadratic splines. To illustrate the effect of

variable selection on the total estimate, figure 1 plots the bias, log of variance, and log of mean squared error (MSE) of the total estimates from each candidate model during the backward selection processes. The backward selection process should have terminated in a model with five variables; however, the entire subset of models for the backward selection method is plotted. The data is generated from superpopulation four with $\sigma_0 = 0.4$ for sample size $n = 200$. Figure 1c clearly shows that the backward selection yields a model with the smallest MSE of the total estimate. Therefore, by selecting the correct set of auxiliary variables, the proposed variable selection procedure not only results in a parsimonious model for auxiliary variables but also improves the accuracy of the total estimates. Furthermore, figure 1a and b show that the selection of auxiliary variables primarily affects the variances of the total estimates more than the bias.

For each combination of noise level and sample size, 100 replicated SRS samples were drawn from the same population. The results of the number of correct fitting models in 100 replications for both forward and backward approaches in the linear and quadratic models are summarized and compared to the simulation results from Wang and Wang (2011) in table 1. A correct fitting model is defined as selecting all of the correct auxiliary variables and none of the incorrect ones, as defined in the superpopulation model.

For all four superpopulation models and two noise levels, the percentage of correct fitting models increases to 100 percent as the sample size increases, as predicted by theorem one. Our method identifies the correct variables more often than the SBLL method, especially at smaller sample sizes (table 1). This can be seen for both the linear and quadratic splines, indicating the linear and quadratic choices for p do not influence the results.

The bias and standard error of the total estimate for each model are compared to the oracle model and the full model using linear splines in table 2. Here, the oracle model contains only relevant auxiliary variables, while the full model contains all ten auxiliary variables. The Horvitz-Thompson estimator, equivalent to using the null model, is also presented in table 2. The number of replications was increased to one thousand in order to obtain stable bias and variance estimates with minimal Monte Carlo error.

Table 2 shows that the bias and standard error of the total estimate for each model decrease as the sample size increases. The bias and standard error using the selected model are almost identical to using the oracle model and smaller than the full model for most comparisons. When the sample size is small ($n = 50$), the selected model can reduce the standard error by more than 30 percent compared with the full model. The selected model, except for superpopulation four, achieves a much lower variance than the Horvitz-Thompson estimator. For example, for sample size $n = 200$, the selected model reduced the standard error of the estimate by 71 percent compared with the

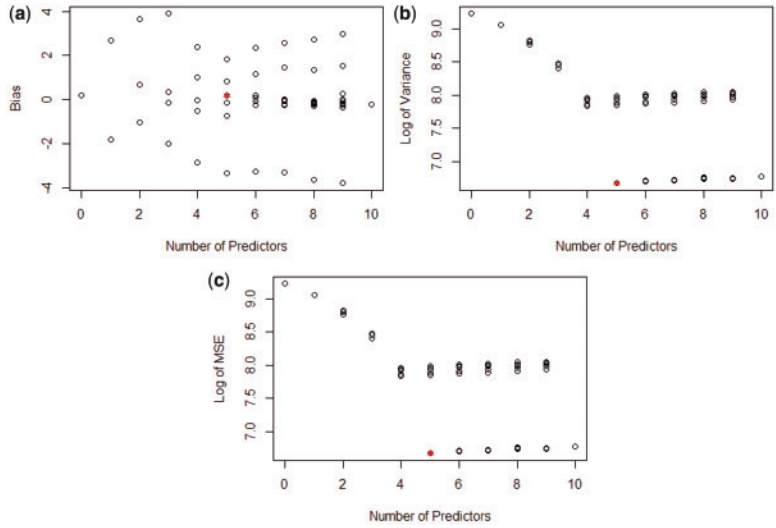

Figure 1. The Bias, Log of Variance, and Log of Mean Squared Error of the Total, Shown for Each Candidate Model During the Backward Selection Processes Under Superpopulation Four with $\sigma_0 = 0.4$ for Sample Size $n = 200$. The pattern is similar for the other models. A) Bias. B) Log of variance. C) Log of mean squared error.

Horvitz-Thompson estimator. We also calculated the bias and standard error of the total estimate using quadratic splines. The results are not presented here since they are similar to the results with linear splines. Here, we focused on comparing the total estimates using the proposed method with those using the full and oracle additive models and the Horvitz-Thompson estimator. Note that the Horvitz-Thompson estimator can be viewed as the null model with no additive component. It is also of interest to compare the proposed method with the generalized regression (GREG) estimator, which is another choice for survey practitioners. The proposed method with an additive model is expected to do better when the relationship between the auxiliary variables and the response are nonlinear. However, the improvement using our method is expected to be less dramatic as compared with the improvement over the Horvitz-Thompson approach.

5. APPLICATION

To illustrate our procedure, we consider the 1999 to 2000 Academic Performance Index (API) growth data set available in the R survey package (Lumley 2014), which tracks changes in academic performance and growth of

California schools with at least one hundred students (see API 2000). Information on the proportion of subsidized school lunches and English language learners, parent education level, and enrollment are included for these schools.

The data set contains a population of 6,194 California schools. In order to illustrate our method and create a complete data set, we eliminated variables with missing data. After running a correlation analysis, variables that were highly collinear were also eliminated. Categorical variables were excluded from this analysis to focus our attention on the relationships that could be estimated using splines. Alternatively, one can include these categorical variables as linear terms and focus on variable selection of nonlinear terms using the BIC in (5) without penalizing the number of parameters for the linear terms. However, it requires further investigation to define an appropriate BIC to perform variable selection for both linear and nonlinear terms simultaneously. After these considerations, nine auxiliary variables remained and are described in table 3. Our goal is to estimate the average API in year 2000 (api00) for the population based on a stratified sample and to select the significant predictors for API using the auxiliary variables in the data set. The API is calculated by the California Department of Education based on a standardized testing of students. The population average is estimated using additive model-assisted estimation based on the forward and backward selection methods. The Horvitz-Thompson estimate and the additive model-assisted estimate based on the full model are calculated for comparison. Note that the proposed model-assisted approach is not feasible for surveys when auxiliary variables cannot be completely observed for unsampled schools.

As in the simulations, one thousand replication samples of size $n = 50$, 100, and 200 were drawn from the population using stratified random sampling. For this illustration, the sample size of each strata was selected by nonproportional allocation: 50 percent to elementary schools, 30 percent to middle schools, and 20 percent to high schools, resulting in unequal selection probabilities. The percentages were chosen to select more schools from the larger strata.

Figure 2 summarizes results for our variable selection methods. At sample size $n = 200$, the variables most often selected were the number of students enrolled (enroll), the percentage of students eligible for subsidized meals (meals), and the percentage of parents with graduate school level education (grd.sch). The known noise variables included in the analysis, the school identifier (cds), and the district number (dnum) and were not selected for most models. The percentage of students who were in their first year (mobil) was excluded from most models, as well. Using both forward and backward selection, about 10–25 percent of models at $n = 200$ included the percentage of parents that are high school graduates or have some college (hsg.col), the percentage of parents that are college graduates (col.grd), and the percentage of English language learners (eng.ll). The average model size and its standard error can be found

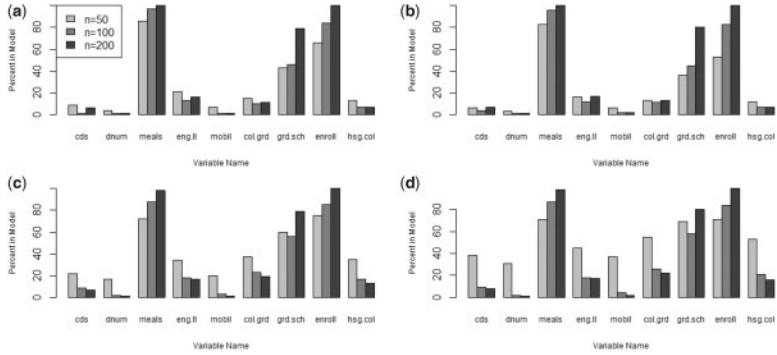

Figure 2. The Percentage of Selected Models Containing Each Auxiliary Variable in the API Data Set. A) Forward selection, linear splines. B) Forward selection, quadratic splines. C) Backward selection, linear splines. D) Backward selection, quadratic splines.

in table 4 for both forward and backward selection methods. The forward selection models had less variables selected than the backward selection at sample size $n = 50$. At sample size $n = 200$, the average model size was about three for both the forward and backward selection methods. The standard error of the model size decreases as sample size increases.

The bias and standard error from estimating the mean API for the population using the forward and backward selection process is presented in table 5. Models selected from both forward and backward selection have lower standard errors than using either the full model or the Horvitz-Thompson estimator. The bias for the backward, forward, and full models ($n = 100$ and $n = 200$) have bias values nearly identical to the bias values for the Horvitz-Thompson estimator, which we know are unbiased estimators. However, for $n = 50$, there was a negative bias for all models.

Models resulting from both the forward and backward methods reduced the standard error of the total estimate compared to the Horvitz-Thompson estimator and the full model with negligible bias. It successfully ruled out known noise variables from the final model for more than 95 percent of simulations with larger sample sizes, demonstrating the effectiveness of our approach.

To explore possible interactions among these variables, we considered interaction terms of three variables (enroll, meals, and grd.sch) that were most relevant, as shown in figure 2. We conducted both forward and backward selection procedures with these three variables in the model and added (or deleted) the interaction terms one at a time. Figure 3 plots the percentage of times that any interaction term is selected in the models. Less than 10 percent of the simulations included any interactions in the models, a finding that supports the additive models we used in this analysis.

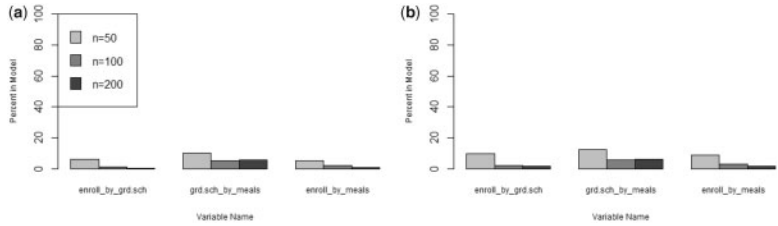

Figure 3. The Percentage of Selected Models Containing Each Interaction Terms in the API Data Set. A) Forward selection, linear splines. B) Backward selection, linear splines.

6. CONCLUSION

The extension of the BIC to additive model-assisted estimation provides a new tool to improve estimates of finite population quantities. The additive model captures the unknown nonlinear relationship, while the BIC reduces the chance of spurious results. The proposed BIC provides a consistent method of variable selection for the purposes of model building.

The size of the parameter space when using spline estimates with the superpopulation model and complex sampling design creates a challenging theoretical problem. The consistency proof of the proposed BIC provides the understanding of its large sample properties. The simulations provide confirmation that the BIC is effective in variable selection at smaller sample sizes. Our method applied to the API data set demonstrates its usefulness for an applied problem. The validity of the proposed BIC method relies on the additive model assumption. The performance of the proposed method is completely unknown when the additive model is misspecified.

Similar to the assumptions adopted in the polynomial spline literature, we have studied the asymptotic properties of the proposed spline estimator when the choice of knots are nonstochastic. In practice, however, the knot sequence is often selected by data driven methods. Therefore, it is of interest to investigate the effect of data-driven knots on the asymptotic results. We leave this to future research.

Future research may improve this method by incorporating a theoretically justified penalty to the likelihood based on the effective sample size resulting from the sampling design. Lumley and Scott (2015) suggest an adjustment based on a design effect for linear models that could be investigated to determine how to adapt their method to nonparametric models. Additional models that vary across strata or clusters can be examined to determine how to account for these in a variable selection method.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Academic Performance Index (2000), available at http://www.cde.ca.gov/ta/ac/ap, Accessed: 2015-12-01.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Breidt, F., G. Claeskens, and J. Opsomer (2005), "Model-Assisted Estimation for Complex Surveys Using Penalised Splines," *Biometrika*, 92, 831–846.

Breidt, F. J., and J. D. Opsomer (2000), "Local Polynomial Regression Estimators in Survey Sampling," *Annals of Statistics*, 28, 1026–1053.

Chen, R., and R. S. Tsay (1993), "Nonlinear Additive ARX Models," *Journal of the American Statistical Association*, 88, 955–967.

De Boor, C. (2001), *A Practical Guide to Splines*, Springer-Verlag New York: Springer.

Dol, W., T. Steerneman, and T. Wansbeek (1996), "Matrix Algebra and Sampling Theory: The Case of the Horvitz-Thompson Estimator," *Linear Algebra and Its Applications*, 237–238, 225–238.

Dorfman, A. H. (1992), "Nonparametric Regression for Estimating Totals in Finite Populations," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 622–625. Alexandria, VA: American Statistical Association.

Fabrizi, E., and P. Lahiri (2007), "A Design-Based Approximation to the BIC in Finite Population Sampling," Technical Report 4, Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Universita degli Studi di Bergamo.

Godambe, V., and V. Joshi (1965), "Admissibility and Bayes Estimation in Sampling Finite Populations," *The Annals of Mathematical Statistics*, 36, 1707–1722.

Hastie, T., and R. Tibshirani (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–310.

Hens, N., M. Aerts, and G. Molenberghs (2006), "Model Selection for Incomplete and Design-Based Samples," *Statistics in Medicine*, 25, 2502–2520.

Huang, J. Z. (1998), "Projection Estimation in Multiple Regression with Application to Functional ANOVA Models," *The Annals of Statistics*, 26, 242–272.

Huang, J. Z., J. L. Horowitz, and F. Wei (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282.

Huang, J. Z., and L. Yang (2004), "Identification of Non-Linear Additive Autoregressive Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 463–477.

Krieger, A. M., and D. Pfeffermann (1992), "Maximum Likelihood Estimation from Complex Sample Surveys," *Survey Methodology*, 18, 225–239.

Lumley, T. (2017), *Survey: Analysis of Complex Survey Samples, R Package Version 3.32*. https://cran.r-project.org/web/packages/survey/citation.html (accessed July 1, 2018).

Lumley, T., and A. Scott (2015), "AIC and BIC for Modeling with Complex Survey Data," *Journal of Survey Statistics and Methodology*, 3, 1–18.

Mallows, C. L. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–675.

Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007), "Model-Assisted Estimation of Forest Resources with Generalized Additive Models," *Journal of the American Statistical Association*, 102, 400–409.

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

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705.

Wang, L., and S. Wang (2011), "Nonparametric Additive Model-Assisted Estimation for Survey Data," *Journal of Multivariate Analysis*, 102, 1126–1140.
Xu, C., J. Chen, and H. Mantel (2013), "Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data," *Survey Methodology*, 39, 303–321.
Xue, L. (2009), "Consistent Variable Selection in Additive Models," *Statistica Sinica*, 19, 1281–1296.
Xue, L., and L. Yang (2006), "Additive Coefficient Modeling via Polynomial Spline," *Statistica Sinica*, 16, 14–23.

APPENDICES

A. SKETCH OF PROOF

The notation of $E[\cdot|U]$ and $P(\cdot|U)$ is adopted to denote the sampling design expectation and probability, respectively, by conditioning on the population U. Without conditioning, $E[\cdot]$ and $P(\cdot)$ denote the expectation and probability, respectively, with respect to the joint distribution of the superpopulation model and the sampling design.

For any $M \subset \{1, \ldots, d\}$, let $\mathbb{H}_M$ be the space of all square integrable additive functions for variables x_l, $l \in M$. Let $\mathbb{G}_M$ be the space of additive spline functions with the form

$$g(\mathbf{x}) = g_0 + \sum_{l \in M} g_l(x_l),$$

where g_0 is a constant and g_l is a spline function with degree p with J_n interior knots. The resulting dimension of $\mathbb{G}_M$ is $q_M = 1 + r(p + J_n)$, where r is the number of auxiliary variables in M. For the purpose of identifiability, assume $\int_{C_l} g_l(x)dx = 0$, for $l \in \{1, \ldots, d\}$, where C_l is the support of X_l. Without loss of generality, it is assumed that C_l is the unit interval. Similar to Huang (1998), we introduce inner products on $\mathbb{H}_M$ as

$$\langle f, g \rangle = E[f(\mathbf{X})g(\mathbf{X})],$$

$$\langle f, g \rangle_N = \frac{1}{N} \sum_{i \in U} f(\mathbf{X}_i)g(\mathbf{X}_i),$$

$$\langle f, g \rangle_n = \frac{1}{N} \sum_{i \in s} \pi_i^{-1} f(\mathbf{X}_i)g(\mathbf{X}_i).$$

The first and second equations are the theoretical and empirical inner products respectively. The last one, $\langle f, g \rangle_n$, can be interpreted as the Horvitz-Thompson estimator of $\langle f, g \rangle_N$. The corresponding norms are $||f||^2 = \langle f, f \rangle$, $||f||_N^2 = \langle f, f \rangle_N$ and $||f||_n^2 = \langle f, f \rangle_n$. The theoretical inner product is used to

define the orthogonal projection onto $\mathbb{G}_M$ and $\mathbb{H}_M$ as $\text{Proj}_{M,n}$ and Proj_M, respectively. Define

$$m^*_{M,n} = \text{Proj}_{M,n} m \text{ and } m^*_M = \text{Proj}_M m. \tag{8}$$

We first present some preliminary results in the following lemmas. Detailed proofs of some results are presented in the supplementary data online.

Lemma one. Under assumptions (A1)–(A6), one has

$$|N^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - m(\mathbf{x}_i))^2 - \sigma^2| = O_P(N^{-1/2}).$$

Proof. It follows from standard arguments for consistency of the Horvitz-Thompson estimator. Theorem 4.2 of Dol, Steerneman, and Wansbeek (1996) is used to bound the variance of the Horvitz-Thompson estimator. ■

Lemma two. Under assumptions (A1)–(A6), one has

$$N^{-1} \sum_{i \in S} \pi_i^{-1} (m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))^2 = O_P(J_n^{-2p-2} + J_n/N).$$

Proof. It gives the L_2 consistency of the polynomial spline estimator for finite populations. The detailed proof is given in the supplementary data online.

Lemma three. Under assumptions (A1)–(A6), one has

$$N^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - m(\mathbf{x}_i))(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)) = O_P(\sqrt{J_n^{-2p-2} + J_n/N}).$$

Proof. It follows directly from the Cauchy-Schwarz inequality and lemmas one and two. ■

Lemma four. Under assumptions (A1)–(A6), one has $|\hat{N}/N - 1| = O_P(N^{-1/2})$, where $\hat{N} = \sum_{i \in S} \pi_i^{-1}$.

Proof. It follows from standard arguments for consistency of the Horvitz-Thompson estimator. Theorem 4.2 of Dol et al. (1996) is used to bound the variance of the Horvitz-Thompson estimator. ■

Lemma five. Let M_0 be the true model. Under assumptions (A1)-(A6), one has

$$|\text{WMSE}_{M_0} - \sigma^2| = O_P(\sqrt{J_n^{-2p-2} + J_n/N}),$$

where WMSE_{M_0} is defined similarly as (7), but with $\hat{m}_{M_k}$ in place of m_{M_0}.

Proof. The numerator of WMSE_{M_0} can be decomposed as,

$$\begin{aligned} N^{-1}\sum_{i\in S}\pi_i^{-1}\varepsilon_i^2 &= N^{-1}\sum_{i\in S}\pi_i^{-1}(y_i - m_0(\mathbf{x}_i))^2 \\ &\quad + N^{-1}\sum_{i\in S}\pi_i^{-1}(m_0(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i))^2 \\ &\quad + 2N^{-1}\sum_{i\in S}\pi_i^{-1}(y_i - m_0(\mathbf{x}_i))(m_0(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)) \\ &= \mathbf{I} + \mathbf{II} + \mathbf{III}. \end{aligned}$$

From the results of lemmas one, two, and three, and Slutsky's theorem, we have that

$$|\mathbf{I} + \mathbf{II} + \mathbf{III} - \sigma^2| \le |\mathbf{I} - \sigma^2| + |\mathbf{II}| + |\mathbf{III}| = O_P(J_n^{-2p-2} + J_n/N). \quad (9)$$

Combining the results in (9) with lemma four using Slutsky's theorem yields

$$\left|\frac{N^{-1}\sum_{i\in S}\pi_i^{-1}\varepsilon_i^2}{N^{-1}\sum_{i\in S}\pi_i^{-1}} - \sigma^2\right| = O_P(J_n^{-2p-2} + J_n/N).$$

Therefore, WMSE_{M_0} is a consistent estimator of σ^2. ■

Lemma six. Let $\mathbb{GG}_{\{1,\ldots,d\}}$ be the spline space using all available auxiliary variables, then under (A1)–(A6), one has $\sup_{g\in\mathbb{G}}|\frac{||g||_N}{||g||} - 1| = o_P(1)$.

Proof. It follows similar as the proof of lemma A.3 in Xue and Yang (2006). The detailed proof is given in the supplementary data online. ■

Lemma seven. Under assumptions (A1)–(A6), $\sup_{g\in\mathbb{G}}|\frac{||g||_n}{||g||} - 1| = o_P(1)$.

Proof. It is the sample analog of lemma six. It follows from lemma six and consistency of the Horvitz-Thompson estimators. ■

Consider a set of variables x_l, $l = 1,\ldots,d$, which contain all relevant auxiliary variables and possibly other irrelevant information. Let $M \subset \{1,\ldots,d\}$ represent a model containing x_l, $l \in M$. Then for $m^*_{M,n}$ and m^*_M defined in (8), we have the following results.

Lemma eight. Under assumptions (A1)–(A6), $||\hat{m}_M - m^*_{M,n}|| = O_P\left(\sqrt{J_n^{-2p-2} + J_n/N}\right)$.

Proof. It is obtained by applying lemma seven to the population version given in Huang (1998). ■

Lemma nine. Under assumptions (A1)–(A6), if M underfits then $c(M, m) = ||m_M^* - m|| > 0$.

Proof. It is obtained by applying lemma seven to the population version given in Huang (1998). ■

Proof of theorem one. The proof closely follows the one given in Huang and Yang (2004). The detailed proof is given in the supplementary data online. ■

B. TABLES

Table 1. Percent of Correct Fitting Models Using Variable Selection in Four Fixed Populations of Size $N=1000$

			Percent correct fits					
			Linear spline		Quadratic spline		SBLL	
Model	σ_0	n	Forward	Backward	Forward	Backward	Forward	Backward
1	0.1	50	98	98	98	98	72	73
		100	99	99	99	99	97	97
		200	100	100	100	100	99	99
	0.4	50	90	89	94	92	76	77
		100	98	98	97	97	98	98
		200	100	100	99	99	100	100
2	0.1	50	97	93	99	97	87	87
		100	100	100	99	99	96	96
		200	100	100	100	100	100	100
	0.4	50	95	91	95	92	79	80
		100	100	100	99	99	98	98
		200	100	100	99	99	100	100
3	0.1	50	97	92	97	95	87	86
		100	97	97	99	99	91	91
		200	98	98	100	100	100	100
	0.4	50	89	82	86	82	83	83
		100	99	99	98	98	99	99
		200	99	99	100	100	100	100
4	0.1	50	81	91	90	95	68	69
		100	97	97	100	100	88	88
		200	99	99	100	100	100	100
	0.4	50	84	91	84	90	69	69
		100	98	98	97	97	97	97
		200	99	99	100	100	100	100

NOTE.—The simulation drew one hundred simple random samples of size n and selected the variables for both forward and backward approaches using the proposed method. The SBLL column presents the results from Wang and Wang (2011) for comparison.

Table 2. Monte Carlo Bias and Standard Error of the Linear Spline Model-Assisted Estimators in Four Fixed Populations of Size $N = 1{,}000$

			Forward		Backward		Oracle		Full		HT	
Model	σ_0	n	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1	0.1	50	0.46	15.32	0.44	15.56	0.49	15.25	−0.38	23.40	−7.32	176.51
		100	0.07	10.12	0.07	10.12	0.03	10.11	−0.25	12.09	−3.43	125.18
		200	0.06	6.84	0.06	6.84	0.07	6.84	0.05	7.30	−0.38	85.01
	0.4	50	1.62	61.38	0.98	64.02	1.96	61.02	−1.51	93.59	−5.59	182.93
		100	0.20	40.60	0.20	40.60	0.10	40.44	−0.99	48.37	−3.48	131.53
		200	0.25	27.35	0.25	27.35	0.26	27.35	0.21	29.18	−0.12	87.79
2	0.1	50	−1.27	43.55	−1.15	44.32	−0.95	43.26	−1.38	65.42	8.15	264.45
		100	−1.67	29.78	−1.65	29.78	−1.67	29.73	−1.65	35.48	−5.35	185.38
		200	−0.36	18.83	−0.36	18.83	−0.40	18.79	−0.53	20.09	2.05	126.64
	0.4	50	−0.67	74.59	−0.14	77.58	−0.34	74.06	−2.51	112.33	9.88	272.16
		100	−2.27	49.39	−2.28	49.39	−2.15	49.26	−2.40	59.86	−5.40	190.63
		200	−0.16	32.70	−0.16	32.70	−0.20	32.69	−0.38	34.93	2.31	129.64
3	0.1	50	−2.95	46.93	−1.68	27.77	−1.61	26.75	−1.75	39.96	10.04	156.50
		100	−1.33	18.11	−1.33	18.11	−1.30	18.10	−1.68	21.35	−0.45	112.00
		200	−0.31	11.94	−0.31	11.94	−0.32	11.94	−0.13	12.48	0.32	74.53
	0.4	50	−2.61	80.98	−0.07	68.60	−0.53	65.97	−2.89	99.57	11.77	165.60
		100	−1.84	44.00	−1.75	44.15	−1.66	44.00	−2.42	52.04	−0.50	118.92
		200	−0.11	29.18	−0.11	29.18	−0.11	29.18	0.03	30.66	0.58	78.91
4	0.1	50	9.27	100.19	0.52	51.56	0.19	49.63	0.09	67.23	5.19	210.22
		100	0.13	30.47	0.13	30.47	0.08	30.38	0.50	33.22	−5.54	143.88
		200	−0.05	19.59	−0.05	19.59	−0.06	19.60	−0.12	20.62	−1.56	98.69
	0.4	50	8.38	122.98	1.18	75.99	1.14	73.22	−0.34	100.20	6.65	216.92
		100	−0.24	45.93	−0.20	45.85	−0.39	45.78	0.07	50.90	−5.49	147.91
		200	0.05	29.18	0.05	29.18	0.08	29.17	0.09	30.29	−1.30	101.27

NOTE.—The simulation drew one thousand simple random samples of size n and selected the variables for both forward and backward using the proposed method. The "oracle" estimated the total using the correct auxiliary variables. The Horvitz–Thompson (HT) estimator's Monte Carlo bias and variance were included for comparison.

Table 3. Variable Definitions

Variable	Role	Definition
api00	Response	API in 2000
stype	Strata	School type (elementary, middle, high school)
cds	Auxiliary	County/District/School code
dnum	Auxiliary	District number
meals	Auxiliary	Percentage of students in the free or reduced price lunch program
eng.ll	Auxiliary	Percentage of students that are English language learners
mobil	Auxiliary	Percentage of students who first attended school this present year
col.grd	Auxiliary	Percentage of parents with college degree
grd.sch	Auxiliary	Percentage of parents with postgraduate education
enroll	Auxiliary	Number of students enrolled
hsg.col	Auxiliary	Percentage of parents with high school degree or some college

SOURCE.—API (2000).

Table 4. Average Model Size in the API Data from One Thousand Monte Carlo Simulations Using Stratified Sampling

Direction	p	n	Avg size	Std dev
Forward	1	50	2.64	1.14
		100	2.63	0.85
		200	3.18	0.74
	2	50	2.28	1.13
		100	2.63	0.87
		200	3.25	0.79
Backward	1	50	3.71	1.58
		100	2.99	1.05
		200	3.35	0.86
	2	50	4.70	2.28
		100	3.09	1.14
		200	3.43	0.92

Table 5. Bias and Standard Error of the Mean Estimate in the API Data from One Thousand Monte Carlo Simulations Using Stratified Sampling

		Forward		Backward		Full		HT	
p	n	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1	50	−1.44	10.93	−1.56	12.36	−2.03	13.67	−0.53	19.47
	100	−0.87	6.83	−0.75	6.89	−1.01	7.86	−1	13.76
	200	−0.36	4.77	−0.38	4.81	−0.48	4.83	−0.49	10.1
2	50	−2.45	14.23	−2.74	27.95	−4.09	37.32	−0.53	19.47
	100	−1.04	7.37	−1	7.7	−1.08	10.63	−1	13.76
	200	−0.33	4.86	−0.39	4.9	−0.46	5.33	−0.49	10.1