AN ABSTRACT OF THE DISSERTATION OF

Addison James for the degree of Doctor of Philosophy in Statistics presented on August 17, 2016.

Title: Information Criterion for Nonparametric Model-Assisted Survey Estimators

Abstract approved: _____

Virginia M. Lesser Lan Xue

Nonparametric model-assisted estimators have been proposed to improve estimates of finite population parameters. More efficient estimators are obtained when the parametric model is misspecified due to the flexibility of nonparametric models. In this dissertation, we derive information criteria to select appropriate auxiliary variables to use in an additive model-assisted method. By removing irrelevant auxiliary variables, our method reduces model complexity and decreases estimator variance. Our proposed Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) account for the sampling design using the first order inclusion probabilities of each element. We approximate the additive nonparametric components using polynomial splines. We establish that the proposed BIC is asymptotically consistent to select the important explanatory variables in a finite population. This result is confirmed by our numerical study under a range of superpopulation models. Our numerical study shows that the AIC tends to overfit and does not show an increase in performance as the sam-

ple size increases. Using the BIC, our proposed method is easier to implement and theoretically justified compared with a previously proposed method.

Information Criterion for Nonparametric Model-Assisted Survey Estimators

by

Addison James

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 17, 2016
Commencement June 2017

Doctor of Philosophy dissertation of Addison James presented on August 17, 2016.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Addison James, Author

ACKNOWLEDGEMENTS

I would like to express my sincere and overwhelming gratitude to my advisors, Dr. Virginia M. Lesser and Dr. Lan Xue. I could not have made it through my academic endeavor without their patience, persistence, and unwavering support. I am forever indebted for the time and energy they have spent guiding me through this process. I am very thankful for the countless hours Dr. Lesser spent mentoring me through these past four years by helping me find a topic, stay on track, and keeping me focused on the big picture. I also would like to thank her for the opportunity to gain experience working on real data in the Survey Research Center. I cannot thank Dr. Xue enough for always being available to counsel me through the theoretical development of this research and being a primary resource in the technical aspects of the mathematical proof. Moreover, I want to thank my advisors for never giving up on me.

Besides my advisors, I wish to thank my committee members, Dr. Charlotte Wickham and Dr. Yuan Jiang for their insightful questions, encouragement, and flexibility in scheduling meetings. I would also like to thank Dr. Daniel Edge for serving as the Graduate School Council representative for my defense and Dr. Susan Carozza for serving during my preliminary exam.

A big thank you to the professors who instructed me during my time at Oregon State University: Dr. Alix Gitelman, Dr. Sarah Emerson, Dr. Cliff Pereira, Dr. Yanming Di, Dr. Claudio Feuntes, and Dr. Paul Murtaugh. I would also like to acknowledge Dr. Fred Ramsey and Jeff Kollath for igniting my interest in statistics during my undergraduate studies. Furthermore, I would like to thank my friend Dr. Ted K. Taylor for

directing me to the field of statistics and my Masters advisor, Dr. F. Jay Breidt for starting me of my path toward my Ph.D. research in nonparametric survey statistics.

I am thankful to my fellow graduate students who helped me through my classes and provided a wonderful work environment. Thank you to Tim Skalland, Bin Zhou, and Chris Wolf for spending countless hours at the chalkboard with me solving proofs. Also, thank you to Sasha Friedman, Nima Dolatnia, Jianfei Zheng, and Katie Jager for letting me join their study group when I was new to the program.

I am grateful to my parents for their love and financial support through my undergraduate studies. I especially wish to thanks my wife, Leah Miranda, for editing my thesis for grammar, providing endless emotional support, encouragement, motivation, and picking me up more times than I can count. I could not have done it without you.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX TABLES

Information Criterion for Nonparametric Model-Assisted Survey Estimators

1 Introduction

In statistical problems, a sample is assumed to be a collection of realizations from an infinite population. An example of this is rolling a fair die multiple times and recording the outcome. The die can be repeatedly rolled, with no fixed sample size, since the die represents an infinite population. Another common assumption is that observations from each draw are independent and follow an identical distribution. In the die example, each roll does not affect the outcome of the next roll. However, in sampling a finite setting, a subset of elements in a fixed population size is selected to represent the whole population.

1.1 Survey Sampling

The survey sampling paradigm assumes a sample is a collection of elements drawn randomly from a fixed finite population. The population consists of elements, each having a fixed value or a set of values. In a simple random sample, each element has the same chance of selection. However, elements of the population can be randomly selected with varying probabilities usually due to selecting units in various stages or phases of sampling. Survey sampling methodology was developed to properly analyze and draw valid inference to account for how the sample was selected.

In stratification, elements in a population can be grouped into strata and elements are selected within strata. This can be used to ensure the sample consists of elements taken from known subgroups (or strata) in the population. However, attempts to purposely select elements to represent specific subclasses without a probability design were shown by Neyman (1934) to be unsatisfactory for achieving a representative sample, while stratified random samples were shown to accurately represent the population. Another feature commonly used to reduce costs of surveys across a wide geographic area is multistage sampling. Mahalanobis (2015) gave methods for maximizing the precision of large-scale multistage samples with respect to the unit cost.

Data collected from survey samples are often used to estimate population parameters, such as the mean or total of the population. Horvitz and Thompson (1952) proposed a design-unbiased estimator of the population total and a design-unbiased estimator of its variance. Although the total estimate was widely used, its variance estimator was inefficient and could produce negative estimates under certain conditions. Yates and Gundy (1953) derived a nonnegative design-unbiased variance estimator that was more efficient than the Horvitz-Thompson variance estimator.

Applications of sampling methods and survey methodology emerged over the next 25 years to provide guidance to practitioners conducting surveys. A comprehensive summary of commonly used survey methods was given in Yates (1946), where the author compared the available methods and showed how they relate to each other. Applications of survey methods were found in the economic and social sciences, as well as in agriculture and biology. Later, Yates (1949) gave a more extensive and less mathematical summary of applications for census surveys. Deming (1950) discussed both

sampling theory and practical applications requiring the reader to only know college algebra. Cochran (1953) followed this work with extensive proofs of the sampling theory results. Hansen et al. (1953) provided a two volume book with details of sampling applications in volume 1 and provided theory and proofs in volume 2. A summary of the applications of sampling methods to agriculture was given by Panse and Sukhatme (1954). Sampling techniques, with applications to surveying human subjects, were discussed by Kish (1965).

A shift in the sampling literature started with Brewer (1963) when he considered the fixed population to be the realization of an underlying stochastic processes. This concept was extended to certain linear models by Royall (1970). The theory of model-assisted estimation assumes that the underlying stochastic process is a superpopulation model that describes the distribution of the variable of interest conditioned on the auxiliary variables [Isaki and Fuller (1982)]. When auxiliary information is available for all elements in the population, such as satellite data or government records, a more efficient estimator of population estimates can be achieved through model-assisted estimation [e.g. Cochran (1953), Cassel et al. (1976), Wright (1983)]. The first examples of model-assisted approaches were ratio and regression estimation. The best linear unbiased estimator properties of the ratio estimator were developed in Brewer (1963), including an expression of the conditional variance of the ratio estimator. Estimation techniques, such as ratio and regression estimation, assume a linear relationship between the response and auxiliary variables.

A broader concept called generalized linear regression estimation was introduced to finite population sampling by Cassel et al. (1977). Generalized linear models allowed

for the error structures to take many forms beyond the normal distribution. Robinson and Särndal (1983) established the asymptotic properties of generalized linear models for finite populations. By considering the case where the superpopulation model is a degenerate random variable, Robinson and Särndal (1983) unified the theory of traditional sampling theory with the superpopulation approach. Another connection in sampling theory was made by Deville and Särndal (1992), who showed that generalized regression estimation can be written as a weighted sum using weights that are calibrated for population totals. A summary of these model-assisted developments can found in Särdnal et al. (1992).

Not all relationships between the response and auxiliary variables are linear. Nonparametric models can be used to better approximate nonlinear relationships. These models relax assumptions on the functional form of the relationship which is necessary when the relationship is nonlinear. The nonparametric models provide more robust estimation with respect to model misspecification.

1.2 Nonparametric Modeling

Nonparametric modeling has recently gained popularity due to its ability in detecting relatively complex nonlinear relationships. In literature, a variety of smoothing techniques have been developed to estimate nonparametric functions, which includes kernel based methods such as kernel smoothing, local polynomial smoothing [Fan (1993)], locally weighted scatter plot smoothing [Cleveland (1979)], and methods based on spline approximations such as regression spline [Stone (1985)], smooth-

ing spline [Wahba (1978)], and penalized regression spline [Ruppert et al. (2003)]. All aforementioned smoothing methods are useful in estimating univariate or low dimensional nonparametric functions. However, there are challenges in estimating nonparametric functions with a large number of predictor variables due to the "curse of dimensionality", which refers to the phenomena that available data becomes too sparse in high-dimensions and it requires exponentially more data points to achieve reliable estimation results.

To partly alleviate the "curse of dimensionality", many semi-parametric models have been proposed that specify additional structures on the nonparametric functions. In this paper, we focus on the additive model ([Stone (1985)], [Hastie and Tibshirani (1986)]). The additive model has been used in a variety of applications including mortality in open heart surgery [Parsonnet et al. (1989)], shadow fading in signal strength [Salo et al. (2007)], and genetic modeling [Lo et al. (1993)] . The additive model assumes the contribution of each covariate to be additive, but the form of each contribution is an unspecified univariate function. [Hastie and Tibshirani (1986)] proposed to estimate the functions through backfitting. Cleveland and Devlin (1988) extended local polynomial regression estimation to multivariate functions using the additive model. Other proposed estimation methods for additive model estimation include marginal integration [Linton and Nielsen (1995)], smooth backfitting [Mammen et al. (1999)], regression spline estimators [Stone (1985), Stone (1994)], and spline-backfitted kernel smoothing [Wang and Yang (2007)].

Developments in nonparametric methods have been adapted to survey sampling to improve the finite population estimates. Kuo (1988) introduced the use of nonparamet-

ric methods to estimate distribution functions from survey data. Nonparametric regression was extended to model-assisted estimation by Dorfman (1992) and was also used for estimation finite population distribution functions by Dorfman and Hall (1993). Chambers et al. (1993) proposed a bias robust nonparametric calibration method for estimating population totals. Chambers (1996) used a nonparametric regression bias correction factor to improve case-weighting. Univariate local polynomial regression was extended to model-assisted estimators by Breidt and Opsomer (2000). They proved the total estimate is asymptotically design-unbiased and consistent. Breidt et al. (2005) proposed a class of univariate estimators based on penalized polynomial splines using a data-driven penalty parameter. Multivariate methods were considered by Opsomer et al. (2007) assuming a generalized additive model and by Breidt et al. (2007) assuming a semiparametric model.

Wang and Wang (2011) used the spline-backfitted local linear (SBLL) estimator to estimate additive functions in data collected using a design based sample. The procedure is shown to have high computational speed and efficiency for large data sets with a large number of observations and auxiliary variables. A variable selection method based on asymptotic mean squared error is presented and the authors state that it is design consistent under simple random sampling.

Nonparametric model-assisted estimation is a flexible approach to build a regression model. The shape of the relationship between the auxiliary information with the response variable need not be known. The drawback is that the model may be overfit, resulting in higher sampling variability [Burnham and Anderson (2003)]. Procedures, such as variable selection, can be used to remove unrelevant variables.

1.3 Variable Selection

Variable selection is a statistical procedure used to select the ”best” subset of available variables to be included in the model. It not only reduces model complexity, but also improves prediction accuracy of statistical models. It plays an important role in statistical learning and modeling of high dimensional data arising in many scientific areas. One challenge in variable selection is that the number of candidate models (2^p) is too large that prohibits a full search even with moderate number of variables p. One fundamental approach to perform variable selection is through stepwise deletion or subset selection using criteria such as Marrows' C_p, Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC). The AIC [Akaike (1974)] minimizes the Kullback-Leibler distance between distributions under the candidate model space and the true model. Schwarz (1978) motivated BIC from a Bayesian point of view and it favors the model that is most plausible according the data at hand. Both information criteria balances the goodness fit of a model with its complexity.

The AIC and BIC that were introduced by Akaike (1974) and Schwarz (1978) did not consider data selected by an unequal probability sampling design. In order to produce the best possible model-based estimator, it is necessary to use a variable selection method that accounts for the sampling design. Hens et al. (2006) proposed an AIC for missing observations and single-stage design based samples. An approximation to the BIC for design-based samples was proposed by Fabrizi and Lahari (2007). Xu et al. (2013) gave an alternative BIC for survey sampled data using a non-Bayesian justification. Theoretically justified derivations for both the AIC and BIC are given by Lumley and Scott (2015) for complex design-based samples. They show the asymptotic prop-

erties of these estimators assuming a linear model.

Model selection approaches have been adapted to nonparametric models. Chen and Tsay (1993) extended the idea of best subset regression to additive models for selecting lagged variables in time series models. Shively et al. (1999) used a hierarchical Bayesian approach for variable selection on the additive model and then estimated its functions with model averaging. Huang and Yang (2004) generalized the AIC and BIC to nonparametric models estimated with spline-smoothing. Lin and Zhang (2006) proposed component selection and smoothing operator for model selection on functional ANOVA models. Xue (2009) introduced consistent variable selection for the additive model using penalized polynomial spline. Huang et al. (2010) applied the adaptive LASSO additive models and proved that it is a consistent method of variable selection.

A wide range of research has focused on variable selection for nonparametric models. Other research focused on variable section in design based samples. Wang and Wang (2011) combined these two areas in their recent paper and proposed a BIC for variable selection in additive models with design based samples. Their BIC is based on asymptotic mean squared errors and the asymptotic theory of variable selection to samples using unequal selection probabilities has not been examined.

In this dissertation, we extended the information criterion of Huang and Yang (2004) to samples from finite populations. We proposed a Bayesian information criterion for consistent variable selection in additive model-assisted estimation. Our proposed method is applicable for data generated from a broad range of survey designs. It is challenging to establish the consistency of the proposed variable selection method under the framework of survey sampling. One difficulty arises from the fact

that the nonparametric model was approximated using a finite set of parameters which increased in size as a function of the sample size. Another difficulty is due to the need to account for two sources of variations the probability sampling design and the data generating process from the superpopulation model. The variable selection method proposed by Wang and Wang (2011) was similar to our method since it assumed an additive model and was applicable to data sampled from finite populations. However, it differed from our method in that our BIC was based on the likelihood function rather than asymptotic mean squared error as in Wang and Wang (2011). Furthermore, our method is consistent under more complex designs beyond the simple random sample. Our numerical study demonstrated that our method had better performance than Wang and Wang (2011) for small sample sizes.

Chapter 2 introduces model-assisted estimation, the additive model, and estimation of the model using splines. We derive an AIC and BIC for nonparametric models, evaluate the numerical performance through simulation, and apply the BIC to the Academic Performance Index (API) Growth data set in Chapter 3. The consistency theorem for the proposed BIC is stated and proved in Chapter 4.

2 Additive Model-Assisted Estimation

2.1 Introduction

Model-assisted estimation uses auxiliary information at the estimation stage by considering the finite population as realizations from a superpopulation. This is typically done assuming a linear model. To accurately capture any nonlinear relationships between the auxiliary information and variable of interest, we can instead assume the superpopulation model has a nonparametric form. Estimation of nonparametric models requires special techniques. In this chapter we introduce the additive model for model-assisted estimation and how to estimate the model using polynomial splines.

2.2 The Model

Let $U_N = \{1, ..., N\}$ 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 $\mathbf{x}_i = (x_{i1}, ..., x_{id})'$ 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 the indicator I_i equal 1 if $i \in S$ and 0 otherwise. We will 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 $\pi_{ij,N} = P_{D_N}(i, j \in S) = P_{D_N}(I_i I_j = 1)$. The subscript N will be omitted to simplify the notation.

Our objective is to efficiently and accurately estimate the finite population total

$t_y = \sum_{i \in U} y_i$. The Horvitz-Thompson (HT) estimator [Horvitz and Thompson (1952)],

$$\hat{t}_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i}$$

is unbiased for the population total. The design variance of the HT estimator is

$$\mathrm{Var}(\hat{t}_{HT}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

and its design-unbiased sample estimate is

$$\widehat{\mathrm{Var}}(\hat{t}_{HT}) = \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

An important feature to consider in sampling is the development of sampling weights associated with each sampled element. The sample weight accounts for the number of elements in the population that the sampled element represents. In this dissertation, we consider only the sampling weights in the analysis.

An estimator that takes advantage of the known auxiliary information may produce more design-efficient estimates. Model-assisted estimation uses auxiliary information at the estimation stage by considering $\{y_i\}_{i=1}^N$ as realizations from a superpopulation, ξ, written as

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

where m is true relationship between the variable of interest and the auxiliary variables, and $\{\varepsilon_i\}_{i=1}^N$ are independent and identically distributed mean zero. The model-assisted

estimator takes the form,

$$\hat{t}_{MA} = \sum_{i \in U} \hat{m}(\mathbf{x}_i) + \sum_{i \in S} \frac{y_i - \hat{m}(\mathbf{x}_i)}{\pi_i}, \tag{2.1}$$

where $\hat{m}$ is an estimate of m using the available sample [Särdnal et al. (1992)].

To accurately capture any nonlinear relationships between the auxiliary information and variable of interest, we will assume m to be of an additive form [Hastie and Tibshirani (1986)], written as

$$m(\mathbf{X}) = \alpha_0 + \sum_{l=1}^{d} \alpha_l(X_l), \tag{2.2}$$

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 $l = 1,...,d$. The additive model can improve the efficiency of the total estimator because its flexibility makes it more robust to model misspecification.

A major challenge in estimating nonparametric functions with more than one variable is dealing with the slow convergence rate, commonly referred to as the "curse of dimensionality." However, the additive structure in model (2.2) allows the estimation of the additive model at the same optimal rate of convergence as the univariate case (Stone, 1985).

2.3 Polynomial Splines

Our goal is to estimate the unknown functions of the additive model using polynomial splines. We define the following notation to denote these estimates. Let $C^p([0,1])$ be the space of p-times continuously differentiable functions. For each auxiliary variable $l = 1,...,d$, define a knot sequence $\kappa_{ln} = \{k_{l0} = 0 < k_{l1} < \cdots < k_{lJ_n} < k_{lJ_n+1} = 1\}$, where J_n is the number of interior knots, for some integer $p > 0$. 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 $\left[k_{l(i-1)}, k_{li}\right), i = 1,...,J_n$, 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 $[0,1]$. With an appropriate choice of knots, such polynomial splines often provide accurate approximations of smooth functions and have better convergence rates than regular polynomials without knots [see De Boor (1978) p. 149].

For a fixed p and J_n, let

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

be the degree p truncated power basis for the spline space ϕ_l with J_n knots, and $(x)_+ = x$ if $x > 0$, else $(x)_+ = 0$. Let $\Gamma_{lj}^*(X_l)$ be the jth element of the vector $\Gamma_l^*(X_l)$. Define the centered basis $\Gamma_l(X_l) = (\Gamma_{l1}(X_l), ..., \Gamma_{l(J_n+p)}(X_l))'$, where $\Gamma_{lj}(X_{li}) = \Gamma_{lj}^*(X_{li}) - N^{-1}\sum_{i\in S}\pi_i^{-1}\Gamma_{lj}^*(X_{li})$. The centered basis for all d variables $\mathbf{X} = \{X_1, ..., X_d\}$ is then,

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

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

$$\theta = (\theta_0, \theta_1', ..., \theta_d')' \tag{2.3}$$

can then be estimated simultaneously using least squares.

Suppose $\mathbf{y} = \{y_1, ...y_N\}$ is known, then the population estimate of θ is

$$\tilde{\theta} = \text{argmin}_\theta \sum_{i=1}^{N} \left(y_i - \theta_0 - \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \theta_{lj}\Gamma_{lj}(x_{li}) \right)^2, \tag{2.4}$$

and

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

where $\mathbf{\Gamma} = (\mathbf{1}, \Gamma(\mathbf{x}_1), ..., \Gamma(\mathbf{x}_N))'$ is the design matrix for the truncated power basis using the entire population. For a fixed $\mathbf{x}$, the population based estimate of m is given as,

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

Since only the sampled values of the variable of interest are observed, $\mathbf{y}_S = (y_i, i \in$

$S)'$, then an appropriate sample estimate of θ is

$$\hat{\theta} = \text{argmin}_{\theta} N^{-1} \sum_{i \in S} \pi_i^{-1} \left(y_i - \theta_0 - \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \theta_{lj} \Gamma_{lj}(x_{li}) \right)^2, \tag{2.5}$$

and

$$\hat{\theta} = [\mathbf{\Gamma}_S' \mathbf{\Pi}_S^{-1} \mathbf{\Gamma}_S]^{-1} \mathbf{\Gamma}_S' \mathbf{\Pi}_S^{-1} \mathbf{y}_S, \tag{2.6}$$

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

$$\hat{m}(\mathbf{x}) = \hat{\theta}_0 + \sum_{l=1}^{d} \sum_{j=1}^{J_n+p} \hat{\theta}_{lj} \Gamma_{lj}(x_l). \tag{2.7}$$

The model estimated from the sample, $\hat{m}$, in (2.7) can be applied to the model-assisted estimator as given in (2.1). In practice, the population estimate $\tilde{m}$ is not available since we do not observe each element of the population, but serves as the theoretical expected value of $\hat{m}$ under the sampling design when the population is fixed. This notation is useful for understanding the asymptotic properties of the estimator, as discussed in Section 4.3.

2.4 Discussion

Wang and Wang (2011) suggested a similar method for estimating the total using the SBLL estimate of the additive model in (2.1). 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 sec-

ond 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 Särdnal 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 it is consistent under simple random sampling, but no proof was provided.

The two-stage SBLL method has superior properties for estimating the additive components, but is computationally intensive since local polynomial smoothing needs to be conducted on each variable in every model. Our goal focuses on variable selection, rather than estimation. In our research we use only a single step of polynomial spline estimates and reduce the number of computations per model. In the next section, we propose a BIC based on the likelihood function for design-based samples with unequal selection probabilities and provide theoretical justification for its consistency.

3 Deriving Information Criteria

3.1 Introduction

In this chapter, we derive the AIC and BIC for the additive model under a complex sampling design and present our simulation results. The derivations establish the underlying assumptions and approximation of each information criterion, and their differences become clear. The simulations establish the finite sample performance of the proposed AIC and BIC. Our method is also compared to other literature and brief discussion is provided of adjustments to account for the sampling design.

3.2 Derivation of the AIC

This derivation is based on the work of Lumley and Scott (2015), and loosely follows the derivation given in Burnham and Anderson (2003). Consider a sample S drawn from finite population U using a probability design. The notation $E[\cdot|U]$ is adopted to denote the sampling design expectation by conditioning on the population U. The general $E[\cdot]$ will denote the expectation with respect to the joint distribution of the superpopulation model and the sampling design. Let f be the true model. Consider a candidate model g with parameter vector θ_n. The length of the parameter vector depends on the sample size since we are considering nonparametric models. The

Kullback-Leibler (KL) divergence of f and g is defined as

$$\mathrm{KL}(f,g) = E\left[\log \frac{f(\mathbf{X})}{g(\mathbf{X}|\theta_n)}\right] = E[\log f(\mathbf{X})] - E[\log g(\mathbf{X}|\theta_n)].$$

A finite population sampling approach is to consider the KL divergence in the population, defined as

$$\mathrm{KL}_U(f,g) = \frac{1}{N}\sum_{i\in U} \log f(\mathbf{X}_i) - \frac{1}{N}\sum_{i\in U} \log g(\mathbf{X}_i|\theta_n). \tag{3.1}$$

We write the population log-likelihood of θ_n with respect to g as

$$\ell(\theta_n) = \frac{1}{N}\sum_{i\in U} \log g(\mathbf{X}_i|\theta_n),$$

where π_i is the first order inclusion probability of element i. The quantity $\ell(\theta_n)$ is a population mean, so it can be estimated from the sample using the Horvitz-Thompson (HT) estimator,

$$\hat{\ell}(\theta_n) = \frac{1}{N}\sum_{i\in S} \pi_i^{-1} \log g(\mathbf{X}_i|\theta_n).$$

However, θ_n is also unknown. We can estimate θ_n with $\hat{\theta}_n = \arg\max_{\theta_n} \hat{\ell}(\theta_n)$, the weighted maximum likelihood estimator.

Our goal is to the select the model that minimizes the KL divergence of f and g. Since f does not depend on θ_n, we ignore the first term in (3.1) and focus on maximizing the population log-likelihood in expectation.

Let

$$\ell(\theta_n) = \frac{1}{N}\sum_{i\in U}\ell_i(\theta_n)$$

and define it's naive estimator as,

$$\hat{\ell}(\hat{\theta}_n) = \frac{1}{N}\sum_{i\in S}\pi_i^{-1}\ell_i(\hat{\theta}_n),$$

where $\ell_i(\theta_n) = \log g(\mathbf{X}_i|\theta_n)$.

At this point, we will omit the n subscript from the parameter vector to make the notation less cumbersome. Consider the Taylor expansion of $\hat{\ell}(\hat{\theta}_n)$ around $\theta_0 = \arg\min_{\theta_n} \mathrm{KL}_U(f,g)$,

$$\begin{aligned}\hat{\ell}(\hat{\theta}_n) &\approx \frac{1}{N}\sum_{i\in S}\pi_i^{-1}\ell_i(\theta_0) + \frac{1}{N}\sum_{i\in S}\pi_i^{-1}\left[\frac{\partial\ell_i(\theta_0)}{\partial\theta_n}\right](\hat{\theta}_n-\theta_0)\\ &\quad + \frac{1}{2N}\sum_{i\in S}(\hat{\theta}_n-\theta_0)'\left[\frac{\partial^2\ell_i(\theta_0)}{\partial\theta_n\partial\theta_n'}\right](\hat{\theta}_n-\theta_0).\end{aligned}$$

The expectation of $\hat{\ell}(\hat{\theta}_n)$ is $E[\hat{\ell}(\hat{\theta}_n)] = E[E[\hat{\ell}(\hat{\theta}_n)|U]]$. First consider the conditional expectation,

$$\begin{aligned}E\left[\hat{\ell}(\hat{\theta}_n)|U\right] &\approx E\left[\frac{1}{N}\sum_{i\in S}\pi_i^{-1}\ell_i(\theta_0)\middle|U\right] + E\left[\frac{1}{N}\sum_{i\in S}\pi_i^{-1}\left[\frac{\partial\ell_i(\theta_0)}{\partial\theta_n}\right](\hat{\theta}_n-\theta_0)\middle|U\right]\\ &\quad + E\left[\frac{1}{2N}\sum_{i\in S}\pi_i^{-1}(\hat{\theta}_n-\theta_0)'\left[\frac{\partial^2\ell_i(\theta_0)}{\partial\theta_n\partial\theta_n'}\right](\hat{\theta}_n-\theta_0)\middle|U\right]\\ &= \frac{1}{N}\sum_{i\in U}\ell_i(\theta_0) - \frac{1}{2}(\hat{\theta}_n-\theta_0)'\frac{1}{N}\sum_{i\in U}\left[\frac{\partial^2\ell_i(\theta_0)}{\partial\theta_n\partial\theta_n'}\right](\hat{\theta}_n-\theta_0).\end{aligned}$$

Next we examine the entire term,

$$\begin{aligned} E\left[E\left[\hat{\ell}(\hat{\theta}_n)\middle|U\right]\right] &\approx E\left[\frac{1}{N}\sum_{i\in U}\ell_i(\theta_0)\right] - E\left[\frac{1}{2}(\hat{\theta}_n-\theta_0)'\frac{1}{N}\sum_{i\in U}\left[\frac{\partial^2\ell_i(\theta_0)}{\partial\theta_n\partial\theta_n'}\right](\hat{\theta}_n-\theta_0)\right] \\ &= E\left[\frac{1}{N}\sum_{i\in U}\ell_i(\theta_0)\right] - \frac{1}{2}\mathrm{tr}\left\{I(\theta_0)E\left[(\hat{\theta}_n-\theta_0)(\hat{\theta}_n-\theta_0)'\right]\right\}. \\ &= E\left[\frac{1}{N}\sum_{i\in U}\ell_i(\theta_0)\right] - \frac{1}{2}\mathrm{tr}[I(\theta_0)\Sigma]. \end{aligned} \tag{3.2}$$

The first term of (3.2) cannot be computed from data. However, an approximation is possible by taking a Taylor expansion about $\hat{\theta}_n$,

$$\begin{aligned} \frac{1}{N}\sum_{i\in U}\ell_i(\theta_0) &\approx \frac{1}{N}\sum_{i\in U}\ell_i(\hat{\theta}_n) + \frac{1}{N}\sum_{i\in U}\left[\frac{\partial\ell_i(\hat{\theta}_n)}{\partial\theta_n}\right](\theta_0-\hat{\theta}_n) \\ &\quad + \frac{1}{2N}\sum_{i\in U}(\theta_0-\hat{\theta}_n)'\left[\frac{\partial^2\ell_i(\hat{\theta}_n)}{\partial\theta_n\partial\theta_n'}\right](\theta_0-\hat{\theta}_n) \\ &= \frac{1}{N}\sum_{i\in U}\ell_i(\hat{\theta}_n) - \frac{1}{2}(\theta_0-\hat{\theta}_n)'\hat{I}(\hat{\theta}_n)(\theta_0-\hat{\theta}_n) \end{aligned}$$

Thus, the expectation is

$$\begin{aligned} E\left[\frac{1}{N}\sum_{i\in U}\ell_i(\theta_0)\right] &\approx E\left[\frac{1}{N}\sum_{i\in S}\pi_i^{-1}\ell_i(\hat{\theta}_n)\right] - \frac{1}{2}E\left[(\theta_0-\hat{\theta}_n)'\hat{I}(\hat{\theta}_n)(\theta_0-\hat{\theta}_n)\right] \\ &\approx E[\ell(\hat{\theta}_n)] - \frac{1}{2}\mathrm{tr}\left\{I(\theta_0)E_D[(\theta_0-\hat{\theta}_n)(\theta_0-\hat{\theta}_n)']\right\} \\ &= E[\ell(\hat{\theta}_n)] - \frac{1}{2}\mathrm{tr}[I(\theta_0)\Sigma]. \end{aligned} \tag{3.3}$$

Combining the results of (3.2) and (3.3),

$$E[\hat{\ell}(\hat{\theta}_n)] \approx \ell(\theta_n) - \text{tr}[I(\theta_0)\Sigma].$$

Therefore, an approximately unbiased estimator of $\ell(\theta_n)$ is $\hat{\ell}(\hat{\theta}_n) + \text{tr}[I(\theta_0)\Sigma]$.

For large samples, $I(\theta_0)\Sigma$ is approximately equal to the identity matrix. Multiplying both side by $-2n$ results in the finite population analog for additive models,

$$AIC_D = -2n\hat{\ell}(\hat{\theta}_n) + 2q_n,$$

where q_n is the number of estimated parameters.

If we assume that the superpopulation model is $Y_i = g(\mathbf{X_i}|\theta_n) + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$, then

$$\begin{aligned}
\hat{\ell}(\hat{\theta}_n) &= \frac{1}{N}\sum_{i\in S} \pi_i^{-1} \ell_i(\hat{\theta}_n|y_i) \\
&= -\frac{1}{N}\sum_{i\in S} \pi_i^{-1} \left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log\hat{\sigma}^2 + \frac{1}{2\hat{\sigma}^2}(y_i - \mu(\mathbf{x}_i|\hat{\theta}_n))^2\right] \\
&= \frac{1}{2N}\sum_{i\in S} \pi_i^{-1}(\log 2\pi - \log\hat{\sigma}^2) - \frac{1}{2}
\end{aligned}$$

where $\hat{\sigma}^2 = N^{-1}\sum_{i\in S}\pi_i^{-1}(y_i - g(\mathbf{x}_i|\hat{\theta}_n))^2$.

If we discard the constant terms, then the resulting AIC is

$$AIC^* = \frac{n}{N}\left(\sum_{i\in S}\pi_i^{-1}\right)\log\left(\frac{\sum_{i\in S}\pi_i^{-1}(y_i - g(\mathbf{x}_i|\hat{\theta}_n))^2}{N}\right) + 2q_n.$$

For a more efficient estimator of $\hat{\sigma}^2$, we can use the Hajek estimator, which results in

$$AIC = \frac{n}{N}\left(\sum_{i\in S}\pi_i^{-1}\right)\log\left(\frac{\sum_{i\in S}\pi_i^{-1}(y_i - g(\mathbf{x}_i|\hat{\theta}_n))^2}{\sum_{i\in S}\pi_i^{-1}}\right) + 2q_n.$$

3.3 Derivation of the BIC

Consider a set of candidate models $\{M_k\}$, each with a vector of parameters θ_k of length $q_{k,n} = d_k(J_n + p)$, where $J_n \asymp n^{1/(2p+3)}$ and d_k are the number of auxiliary variables in model M_k. To simplify the notation, k will be omitted from $q_{k,n} = q_n$. Let $p(k)$ be the prior probability for model M_k and $g(\theta_k|k)$ denote the prior of θ_k given model M_k. We assume that the second order derivatives of the likelihood, $L(\theta_k|\mathbf{y})$, exist and are continuous. Using Bayes' Theorem, we can write the joint posterior of the model and parameter vector as

$$f(k,\theta_k|\mathbf{y}) = \frac{p(k)L(\theta_k|\mathbf{y})g(\theta_k|k)}{f(\mathbf{y})}.$$

Integrating over θ_k yields the posterior probability for model M_k,

$$P(k|\mathbf{y}) = \frac{p(k)\int L(\theta_k|\mathbf{y})g(\theta_k|k)d\theta_k}{f(\mathbf{y})}.$$

The objective is to choose the model with the highest posterior probability, which is equivalent to minimizing

$$-2\log P(k|\mathbf{y}) = 2\log f(\mathbf{y}) - 2\log p(k) - 2\log\int L(\theta_k|\mathbf{y})g(\theta_k|k)d\theta_k. \tag{3.4}$$

Since $2\log f(\mathbf{y})$ is constant for all models, it can be discarded. Since we have no prior information on important variables, we will assume that all models are equally likely, $-2\log p(k)$ is discarded as well. Therefore, the final term is the focus of this derivation.

In order to understand the asymptotic behavior of the likelihood, we write the likelihood as,

$$L(\theta_k|\mathbf{y}) = \exp\{\log L(\theta_k|\mathbf{y})\}$$

and take a second order Taylor Expansion about the maximum likelihood estimate, $\hat{\theta}_k$,

$$\begin{aligned}
\log L(\theta_k|\mathbf{y}) &\approx \log L(\hat{\theta}_k|\mathbf{y}) + (\theta_k - \hat{\theta}_k)' \frac{\partial \log L(\theta_k|\mathbf{y})}{\partial \theta_k}\bigg|_{\theta_k=\hat{\theta}_k} \\
&\quad + \frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[\frac{\partial^2 \log L(\theta_k|\mathbf{y})}{\partial \theta_k \partial \theta_k'}\bigg|_{\theta_k=\hat{\theta}_k}\right](\theta_k - \hat{\theta}_k) \\
&= \log L(\hat{\theta}_k|\mathbf{y}) - \frac{1}{2}(\theta_k - \hat{\theta}_k)'\hat{\mathfrak{J}}(\hat{\theta}_k)(\theta_k - \hat{\theta}_k) \qquad (3.5)
\end{aligned}$$

where,

$$\hat{\mathfrak{J}}(\hat{\theta}_k) = \left[-\frac{\partial^2 \log L(\theta_k|\mathbf{y})}{\partial \theta_k \partial \theta_k'}\bigg|_{\theta_k=\hat{\theta}_k}\right]$$

is the observed Fisher information matrix. The linear term disappeared in (3.5) because $\hat{\theta}_k$ maximizes the likelihood.

The sample likelihood can be expressed as,

$$\log L(\theta_k|\mathbf{y}) = n\left[\frac{1}{N}\sum_{i\in S}\pi_i^{-1}\log L(\theta_k|y_i)\right],$$

and hence,

$$\hat{\mathcal{I}}(\hat{\theta}_k) = \left[-n \frac{\partial^2}{\partial \theta_k \partial \theta_k'} \frac{1}{N} \sum_{i \in S} \pi_i^{-1} \log L(\theta_k | y_i) \Big|_{\theta_k = \hat{\theta}_k} \right] = n \bar{\mathcal{I}}(\hat{\theta}_k),$$

where $\bar{\mathcal{I}}(\hat{\theta}_k)$ is the weighted average observed Fisher information.

Since the Horvitz-Thompson estimator is a consistent estimator of population means,

$$\frac{1}{N} \sum_{i \in S} \pi_i^{-1} \log L(\theta_k | y_i) \xrightarrow{p} \frac{1}{N} \sum_{i=1}^{N} \log L(\theta_k | y_i). \tag{3.6}$$

By the Weak Law of Large Numbers,

$$\frac{1}{N} \sum_{i=1}^{N} \log L(\theta_k | y_i) \xrightarrow{p} E[L(\theta_k | y)] \tag{3.7}$$

Combining (3.6) and (3.7),

$$\frac{1}{n} \log L(\theta_k | \mathbf{y}) \xrightarrow{p} E[L(\theta_k | y)]. \tag{3.8}$$

Therefore, by (3.8) $n^{-1} \log L(\theta_k | \mathbf{y})$ is a consistent estimator of $E[L(\theta_k | y)]$ and

$$\frac{1}{n} \hat{\mathcal{I}}(\hat{\theta}_k) \xrightarrow{p} \mathcal{I}(\theta_k),$$

where $\mathcal{I}(\theta_k)$ is the Fisher information in a single observation y.

Next, we focus on the integral from (3.4). Using the Laplace approximation at $\hat{\theta}_k$

and (3.5),

$$\begin{aligned}
\int L(\theta_k|\mathbf{y})g(\theta_k|k)d\theta_k &\approx L(\hat{\theta}_k|\mathbf{y})g(\hat{\theta}_k|k)\int \exp\left[-\frac{1}{2}(\theta_k-\hat{\theta}_k)'\hat{\mathcal{J}}(\hat{\theta}_k)(\theta_k-\hat{\theta}_k)\right]d\theta_k \\
&= L(\hat{\theta}_k|\mathbf{y})g(\hat{\theta}_k|k)\int \exp\left[-\frac{1}{2}(\theta_k-\hat{\theta}_k)'n\bar{\mathcal{J}}(\hat{\theta}_k)(\theta_k-\hat{\theta}_k)\right]d\theta_k \\
&= L(\hat{\theta}_k|\mathbf{y})g(\hat{\theta}_k)(2\pi)^{q_n/2}\left|n\bar{\mathcal{J}}(\hat{\theta}_k)\right|^{-1/2} \\
&= L(\hat{\theta}_k|\mathbf{y})g(\hat{\theta}_k)(2\pi/n)^{q_n/2}\left|\bar{\mathcal{J}}(\hat{\theta}_k)\right|^{-1/2}. \qquad (3.9)
\end{aligned}$$

Substituting this into (3.4) and applying the integral approximation from (3.9),

$$\begin{aligned}
-2\log P(k|\mathbf{y}) &\approx -2\log\left(L(\hat{\theta}_k|\mathbf{y}))g(\hat{\theta}_k)(2\pi/n)^{q_n/2}\left|\bar{\mathcal{J}}(\hat{\theta}_k)\right|^{-1/2}\right) \\
&= -2\log L(\hat{\theta}_k|\mathbf{y}) - 2\log g(\hat{\theta}_k) - q_n\log(2\pi) \\
&\quad + q_n\log n + \log\left|\bar{\mathcal{J}}(\hat{\theta}_k)\right|. \qquad (3.10)
\end{aligned}$$

Dropping terms from (3.10) that do not depend on n,

$$-2\log P(k|\mathbf{y}) \approx -2\log L(\hat{\theta}_k|\mathbf{y}) - q_n\log(2\pi) + q_n\log n.$$

Although $q_n\log(2\pi)$ depends on n, it is asymptotically dominated by $q_n\log n$ so it is discarded. The remaining terms form the nonparametric finite sample analog to the original BIC,

$$\mathrm{BIC}(M_k) = -2\log L(\hat{\theta}_k|\mathbf{y}) + q_n\log n \qquad (3.11)$$

This is the BIC presented in Chapter 3 which is proved to be consistent in Chapter 4.

3.3.1 An Alternative Approach

A correction was proposed by Lumley and Scott (2015) for small samples with a large design effect. We have extended this to our BIC for nonparametric equations. Consider (3.11) without dropping the Fisher information term from (3.10). Let $\tilde{\mathcal{J}}(\theta_k)$ be the Fisher information from one observation if a simple random sample had been drawn. If we add and subtract the logarithm of this term,

$$\begin{aligned} -2\log P(k|\mathbf{y}) &\approx -2\log L(\hat{\theta}_k|\mathbf{y}) + q_n \log n + \log\left|\bar{\mathcal{J}}(\hat{\theta}_k)\right| - \log\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right| + \log\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right| \\ &= -2\log L(\hat{\theta}_k|\mathbf{y}) + q_n \log n + \log\left(\left|\bar{\mathcal{J}}(\hat{\theta}_k)\right|\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right|^{-1}\right) + \log\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right| \\ &= -2\log L(\hat{\theta}_k|\mathbf{y}) + q_n \log n + \log(D) + \log\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right| \end{aligned}$$

where

$$D = \left|\bar{\mathcal{J}}(\hat{\theta}_k)\right|\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right|^{-1}$$

is the design effect.

We drop $\log\left|\tilde{\mathcal{J}}(\hat{\theta}_k)\right|$ since it does not depend on n yielding the design effect correct BIC,

$$\mathrm{BIC_D}(M_k) = -2\log L(\hat{\theta}_k|\mathbf{y}) + q_n \log n + \log(D). \tag{3.12}$$

The design effect, D, goes to zero as the sample size increases, and hence is asymptotically equivalent to our proposed BIC.

3.4 Simulation Results

3.4.1 BIC on Simple Random Sampling

We ran simulations to evaluate the numerical performance of our proposed model selection criterion assuming a finite population. The setup of our simulation is identical to that used by Wang and Wang (2011) so that the variable selection criterion for the two-step SBLL estimator could be directly compared to our proposed selection criterion.

Following Wang and Wang (2011), the following four superpopulation models were considered to generate observations of the population.

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, auxiliary variables, $\mathbf{X}_i = (X_{i,1}, X_{i,2}, ..., X_{i,10})^{\mathrm{T}}$ were independently generated from the [Uniform(0,1)]^{10} distribution for $i = 1, \ldots, N = 1000$. The noise realizations, $\{\varepsilon_i\}_{i=1}^{N}$, were independently simulated from the standard normal distribution. The scale parameter, σ_0, took values 0.1 and 0.4. Simple random samples of size n =50, 100, and 200 were drawn without replacement from the finite population.

For each sample, the proposed variable selection method was applied. The model with the lowest BIC score will be referred to as the *selected model*. The selected model

was then used to estimate the finite population total of interest, t_y, through the additive model-assisted method (2.1). For comparison, we also estimated 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 2.2).

The additive model was estimated using splines. We considered both linear and quadratic splines with knots spaced evenly between 0 and 1. To illustrate how well splines approximate smooth functions, we considered the stronger nonlinear polynomial spline estimates of superpopulation models 3 and 4 using the *oracle model* with $n = 100$. The oracle model estimates the additive model from the data using only the correct variables. Figure 3.1 shows a partial residual plot with spline estimates of $\alpha_5(X_5) = \exp(X_5 - 1)$ from superpopulation model 3 and $\alpha_1(X_1) = \sin(2\pi(X_1 - 0.5))$ from superpopulation model 4. Partial residuals enable the effect of a particular auxiliary variable to be seen after accounting for the others. This is accomplished by subtracting the estimated effect of the other auxiliary variables not in the plot from the response value. For this simulation, the linear spline was chosen to have 2 interior knots and the quadratic spline was chosen to have 1 interior knot. The model selection criterion was applied through forward selection and backward selection that we discuss next.

3.4.1.1 Implementation

In the forward selection process, the initial model fit includes only the intercept and the proposed BIC value is calculated. Next, candidate models are generated by adding

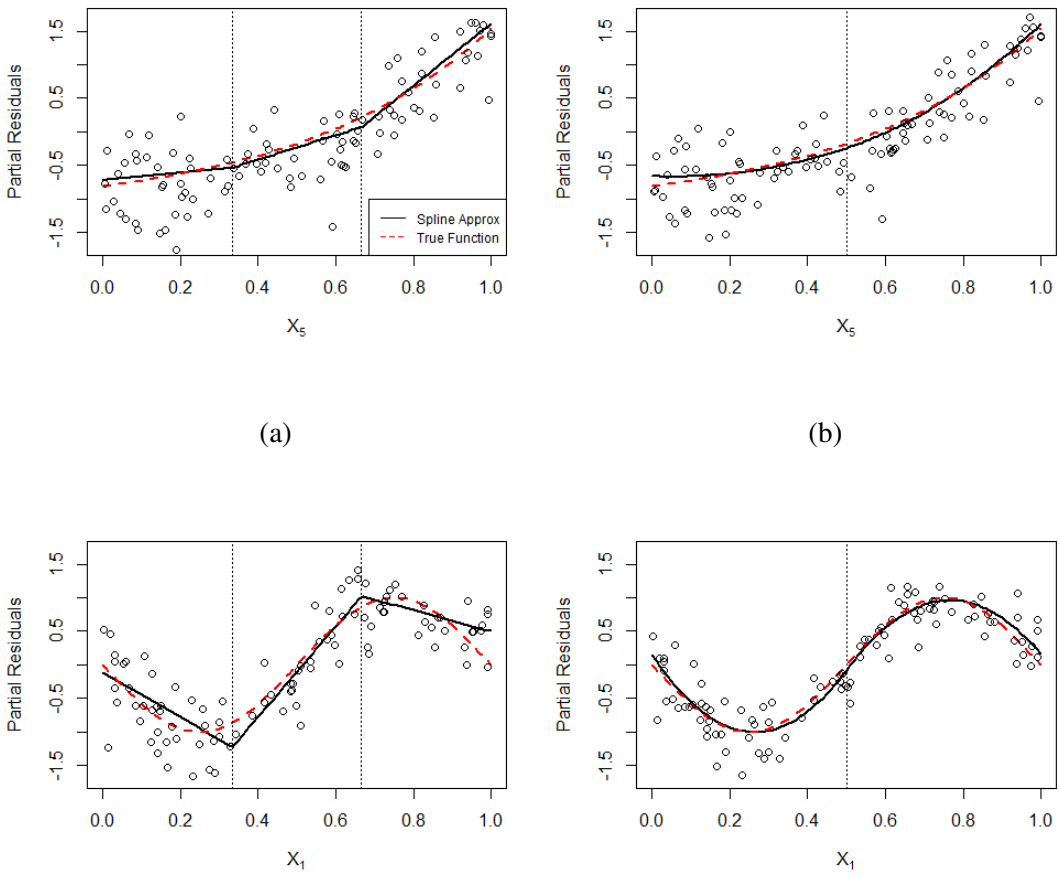
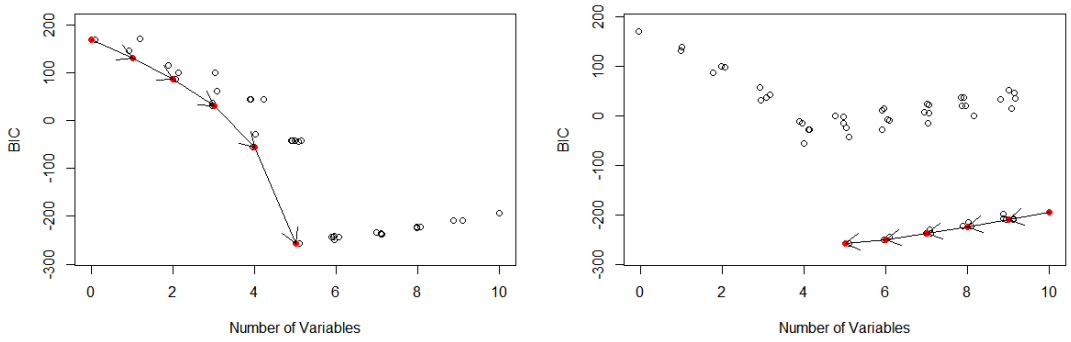


Figure 3.1: Partial residual plots with overlaid spline approximations of the marginal function $\alpha_5(X_5) = \exp(2X_5 - 1)$ under superpopulation model 3 and $\alpha_1(X_1) = \sin(2\pi(X_1 - 0.5))$ under superpopulation model 4 for $n = 100$. The dotted line is the true underlying marginal function and the solid line is the spline estimate under the oracle model. (a) Superpopulation model 3: linear spline. (b) Superpopulation model 3: quadratic spline. (c) Superpopulation model 4: linear spline. (d) Superpopulation model 4: quadratic spline.

each auxiliary variable to the model, one at a time, and the candidate model with the lowest proposed BIC value is retained provided its BIC value is lower than the BIC for initial model. If a candidate model is retained, the process continues adding in the remaining variables one at a time and retaining the model with the lowest BIC value, as seen in Figure 3.2a.



(a) (b)

Figure 3.2: Example of the (a) forward and (b) backward selection process when 5 of 10 auxiliary variables are in the true model. The dots are the BIC value for each model considered. The red dots indicate the lowest BIC value for a particular number of parameters and the arrows indicate the direction of how the process moves from one model to the next.

The backward selection process is the reverse of the forward process.The initial model is the full model including all auxiliary variables. The proposed polynomial spline method is used to estimate the model, and the resulting BIC of the full model is calculated. The candidate models are fit by removing each variable one at a time and the BIC values are calculated. The candidate model with the lowest BIC value is

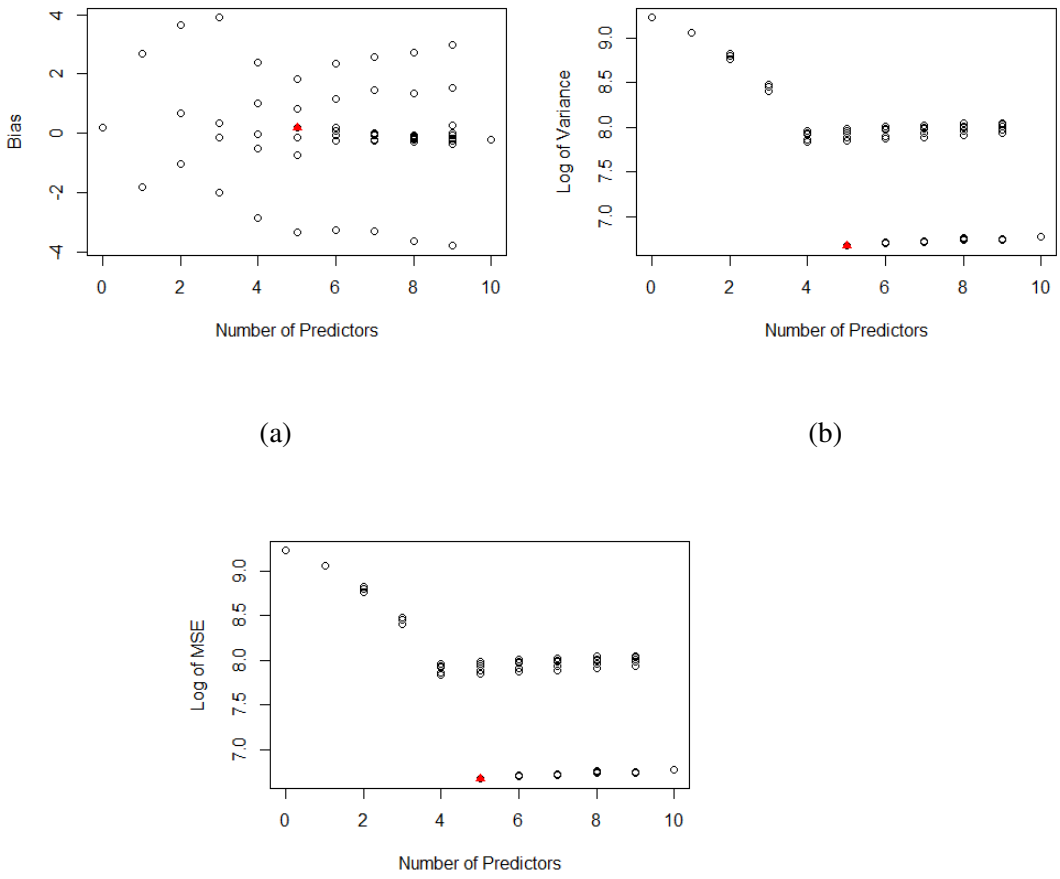
retained, provided it is lower than the current model's BIC. The process of removing variables one at time continues until the current model has a lower BIC than any of the candidate models or there are no variables left in the model. Figure 3.2b illustrates this process.

The variable selection yields a model that, on average, minimizes the mean squared error of total estimate as shown in Figure 3.3c. For example, the bias, variance, and mean squared error are shown for each candidate model in the backward selection processes under superpopulation 4 with $\sigma_0 = 0.4$ for sample size $n = 200$ because it is the most challenging to estimate (Figures 3.3a, 3.3b, and 3.3c, respectively). The pattern is similar for the other models. The backward selection process should have terminated with a model with 5 variables, however, the entire subset of models for the backward selection method is shown.

3.4.1.2 Results

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 A.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. Figure 3.4 provides a graphical comparison of the number of correct fits between the forward and backward methods for Superpopulation Model 4 with $\sigma_0 = 0.4$.

For all four superpopulation models and two noise levels, the percentage of correct



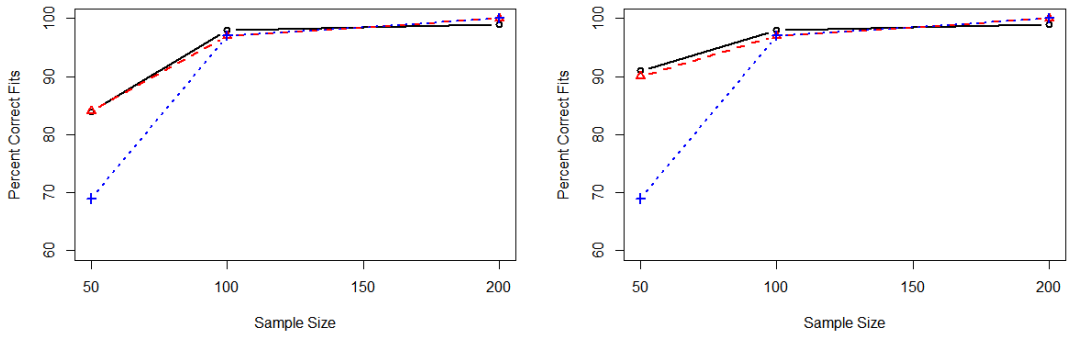
(a) (b)

(c)

Figure 3.3: 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 4 with $\sigma_0 = 0.4$ for sample size $n = 200$ because it is the most challenging to estimate. The pattern is similar for the other models. (a) Bias. (b) Log of Variance. (c) Log of Mean Squared Error.

fitting models increases to 100% as the sample size increases. This is expected for a consistent method. Our method identifies the correct variables more often than the

SBLL method, especially at smaller sample sizes (Table A.1). This can be seen in both the linear and quadratic splines, indicating that linear and quadratic choices for p do not influence the results.

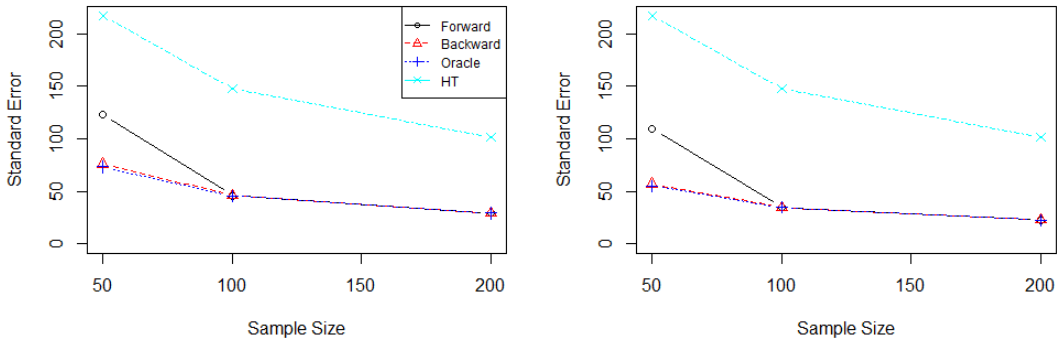


(a) (b)

Figure 3.4: Graphical comparison of the number of correct fits for (a) forward and (b) backward selection between the approaches under superpopulation 4 for $\sigma_0 = 0.4$. The solid and dashed line are the number of correct fits using a linear and quadratic spline, respectively. The dotted line is the number of correct fits using the SBLL method.

The bias and standard error of the total estimate for each model is compared to the oracle model, using linear splines in Table A.2, and quadratic splines in Table A.3. The Horvitz-Thompson estimator, equivalent to using the null model, is also presented in the tables. The number of replications was increased to 1000 in order to obtain bias and variance estimates with minimal Monte Carlo error.

These results show 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 for most comparisons.



(a) (b)

Figure 3.5: Graphical comparison of the standard error between selection method and choice of p under simple random sampling. Both figures are using results from Superpopulation Model 4 with $\sigma_0 = 0.1$. The additive model is estimated with (a) Linear Splines and (b) Quadratic Splines.

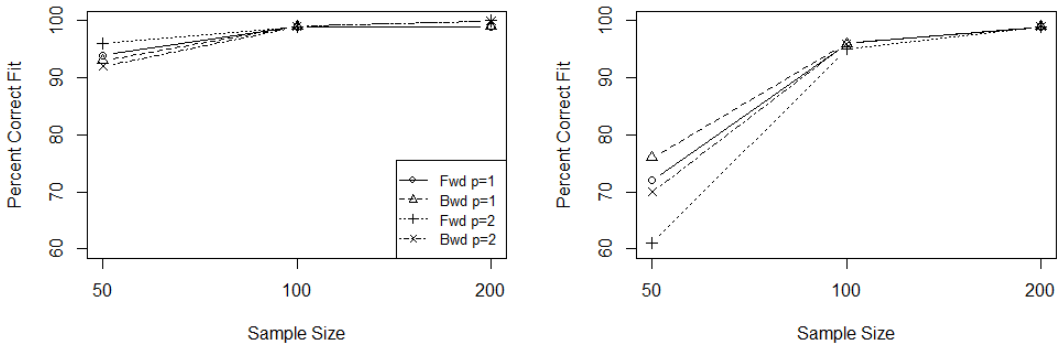
The selected model, except for superpopulation 4, achieves a much lower variance than the Horvitz-Thompson estimator. For example, results from superpopulation model 4 with $\sigma = 0.4$ are presented in Figure 3.5. At sample size $n = 200$ using a linear spline, the selected model reduced the standard error of the estimate by 71% compared to the Horvitz-Thompson estimator. The choice of the degree of the splines, p, did not affect these results.

3.4.2 BIC on Stratified Sampling

We ran simulations assuming a stratified sampling design to demonstrate the performance of our proposed selection method using unequal selection probabilities. Other

than incorporating features of the the sampling design, the setup and implementation of these simulations were the same as those described in section 3.4.1. The population was divided into strata of equal size $N_h = 250, h = 1, ..., 4$. For each strata, a percentage of the total sample size was allocated: 10% to Strata 1, 20% to Strata 2, 30% to Strata 3, and 40% to Strata 4. This created unequal first order inclusion probabilities.

3.4.2.1 Results

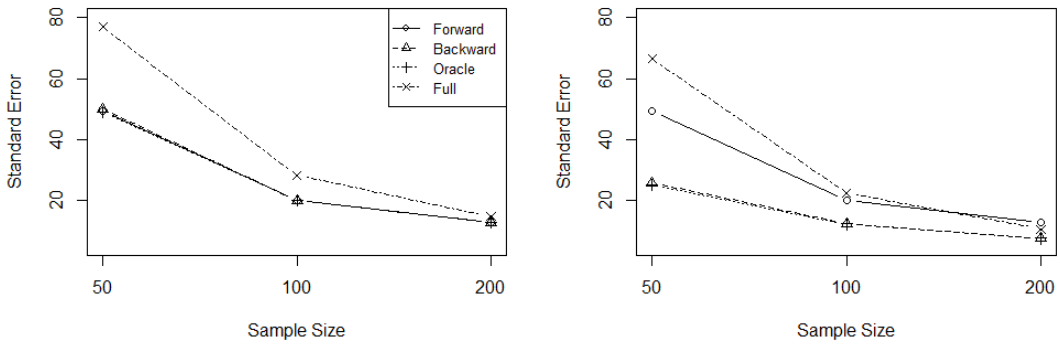


(a) (b)

Figure 3.6: Graphical comparison of the number of correct fits between the selection method and choice of p under stratified sampling. (a) Superpopulation Model 1, $\sigma_0 = 0.1$. (b) Superpopulation Model 4, $\sigma_0 = 0.4$.

The percentage of correct fitting models in 1000 replications is summarized in Table A.4. We see that the percentage of correct fitting models increases to 100% as the sample size increases, demonstrating the consistency of our method under stratified

sampling. This can be seen for both the linear and quadratic splines in Figure 3.6, indicating that the choices for p do not influence the results for stratified samples.



(a) (b)

Figure 3.7: Graphical comparison of the standard error between the selection method and choice of p under stratified sampling. Both figures are using results from Superpopulation 2 with $\sigma_0 = 0.1$. The additive model is estimated with (a) Linear Splines and (b) Quadratic Splines.

The bias and variance of the total estimate for each model is compared to the oracle and full model using linear splines in Table A.5 and quadratic splines in Table A.6. The Horvitz-Thompson estimator is also presented each table.

The results from Tables A.5 and A.6 show that the standard error of the total estimate using the selected model decreases as the sample size increases. Under superpopulation models 1, 2 and 3, the bias decreases with sample size; however, it is unclear from this simulation if the linear spline bias under superpopulation 4 is decreasing. The bias and standard error of the selected and oracle model are almost identical for most comparisons. All models achieve a much lower variance than the Horvitz-Thompson

estimator and the full model. For example, for superpopulation 1 with $\sigma = 0.1$ and sample size $n = 200$ using a linear spline, the selected model reduced the standard error of the estimate by 93% compared to the Horvitz-Thompson estimator and 17% compared to the full model. The choice of the degree of the splines, p, as seen in Figure 3.10, did not affect these results.

3.4.2.2 Alternative Approach Results

We briefly introduced an alternative approach by Lumley and Scott (2015) in Section 3.3.1. We ran simulations under the same setup as described in Section 3.4.2 using the design BIC presented in (3.12), which includes the design effect. The results, presented in Table A.7, showed minimal improvement in the percentage of correct fitting models over our proposed BIC without the design effect given in (4.2). This is due to the increased penalty term in (3.12) and the BIC in (4.2) tended to overfit. The minor improvement in the percentage of correct fits came at the cost of extra calculations in (3.12). At sample size $n = 200$, there was almost no difference in the percentage of correct fitting models between the proposed BIC and the alternative approach. This is expected because they are asymptotically equivalent.

To understand how the alternative approach performs in a situation where our BIC tends to underfit, we consider the following two superpopulation models that contained one variable with a relatively weaker signal.

5. $Y = -1 + 2X_3 + 0.2X_6 + \sigma_0\varepsilon$,

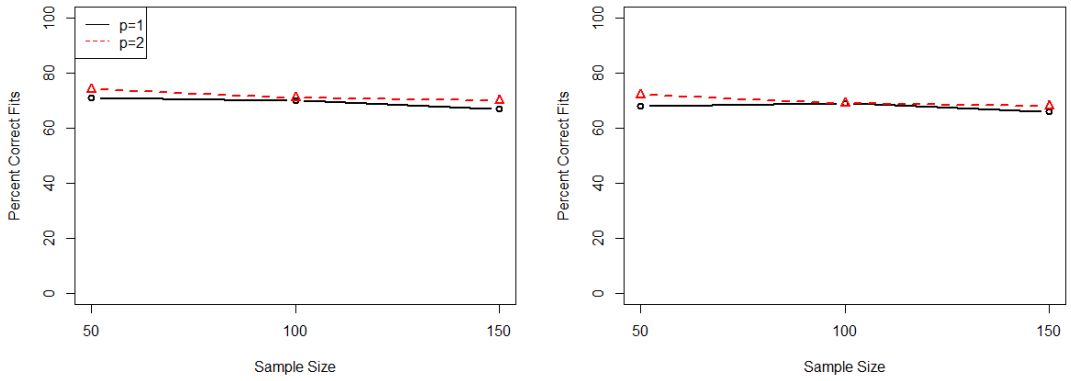
6. $Y = 8(X_2 - .5)^2 + \exp(0.4X_5 - 1) + 2\sin\{2\pi(X_8 - .5)\} + \sigma_0\varepsilon$,

Note that $\alpha_6(X_6)$ in superpopulation model 5 and $\alpha_5(X_5)$ in superpopulation model 6 term have lower signal compared to the other variables in the model. We ran stratified sampling simulations as described in Section 3.3.1 using the two additional superpopulation models. The results are presented in Table A.8. As expected, our proposed BIC in (4.2) tended to underfit. The alternative approach in (3.12) showed decreased performance because it tended to underfit more than (4.2) without the design effect. For example, for superpopulation model 5 at noise level $\sigma_0 = 0.1$ and sample size $n = 50$, our BIC without the design effect underfit in 24% of samples, whereas the BIC with the design effect underfit in 52% correct fits.

The results of this simulation suggest that the design effect proposed by Lumley and Scott (2015) may be unnecessary. Not only does the design effect involve more calculations, it does not yield better results than our proposed method. Furthermore, we cannot theoretically justify the design effect. There may exist a different way to implement a design effect. It would be interesting to justify and account for a design effect in the BIC for future research.

3.4.3 AIC on Stratified Sampling

We ran simulations with the AIC derived in Section 3.2. The same setup was used as Section 3.4.2, including the stratified sampling design, four superpopulation models, two noise levels, and 1000 replications. These simulations give insight into the performance of this AIC when the set of available auxiliary variables include all the important variables.



(a) (b)

Figure 3.8: Graphical comparison of the number of correct fits for (a) forward and (b) backward selection between the approaches and choice of p under superpopulation 1 for $\sigma_0 = 0.4$.

The percentage of correct fitting models in 1000 replications is summarized in Table A.9. For nearly all combinations of superpopulation models, splines, sample sizes, and noise level we examined, the percent of correct fitting models is less than 75% and does not appear to increase as the sample size increases. This can be seen for both both the linear and quadratic splines. Note that the number of correct fits for superpopulation 1 decreases with sample size, as seen in Figure 3.8. Theses results are expected because the AIC is not a consistent variable selection method.

These simulations indicate the BIC derived in Section 3.3 may be a better choice of variable selection method because the AIC will overfit more often. Models containing unimportant variables will have a larger standard error for the total estimate, as demonstrated in Figure 3.3. Therefore, we turn our attention completely to the BIC for the

remainder of this thesis.

3.5 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)]. The API is a measure of California schools' academic performance. It is essential to California's Public Schools Accountability Act of 1999 (see PSAA (2000)) to track the changes in academic performance and growth of each school. Data is available for all California schools with at least 100 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. After these considerations, 9 quantitative auxiliary variables remained and are described in Table 3.1. Our goal is to estimate the average API in 2000 (api00) for the population based on a stratified sample and to select the significant predictors for this 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

Table 3.1: Variable Definitions. Source: API (2000).

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

method described in Section 3.4.1.1. The Horvitz-Thompson estimate and the additive model-assisted estimate based on the full model is calculated for comparison.

Similar to the simulations, 1000 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 non-proportional allocation: 50% to elementary schools, 30% to middle schools, and 20% to high schools. This resulted in unequal selection probabilities due to the stratification. The percentages were chosen such that more elements are selected from larger strata.

Figure 3.9 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

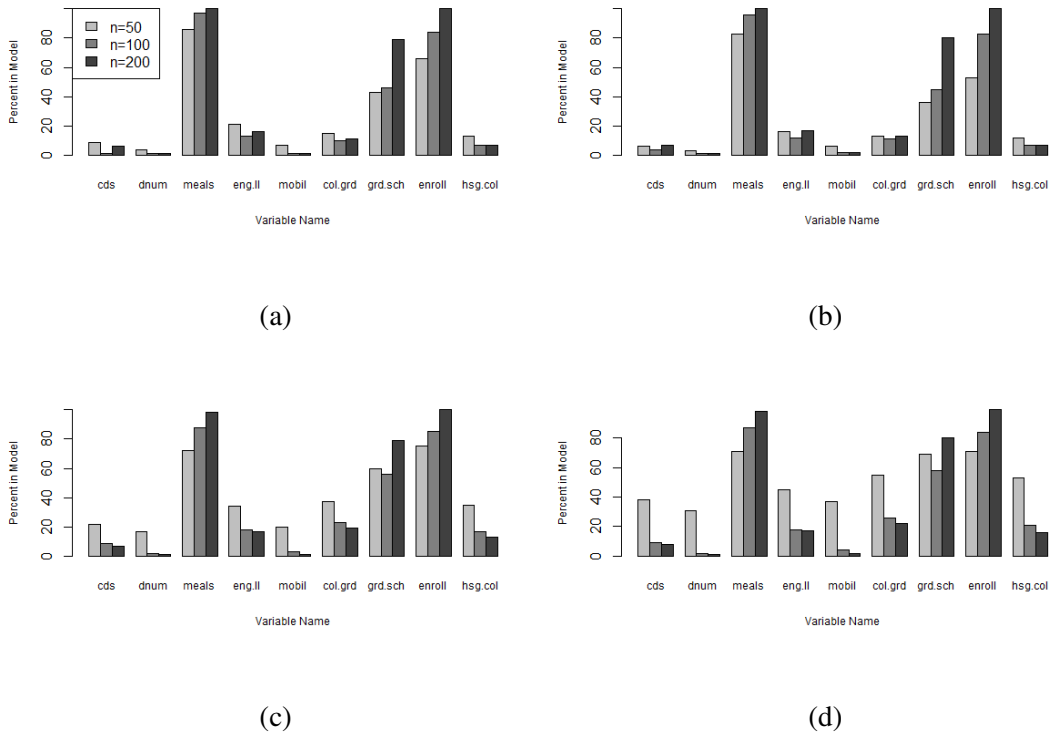
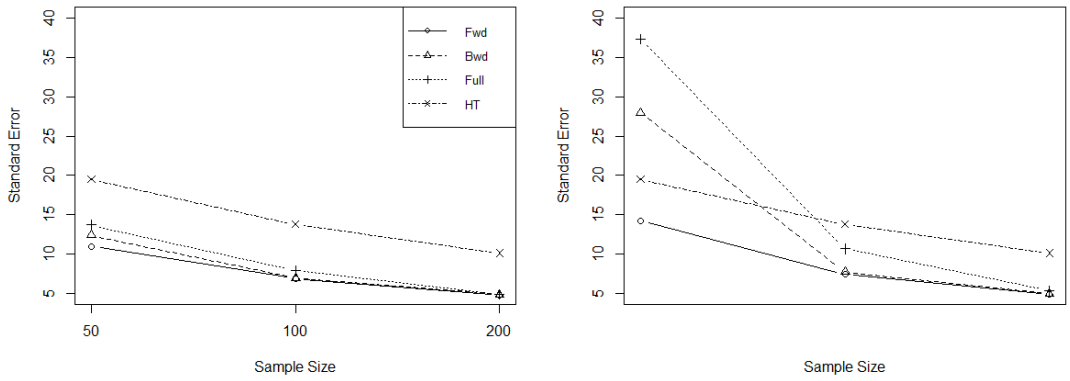


Figure 3.9: The percentage of selected models containing each auxiliary variable in the API data set. (a) Forward Selection, Linear Splines. (b) Forward Selection, Quadratic Splines. (c) Backward Selection, Linear Splines (d) Backward Selection, Quadratic Splines

(dnum) were not selected for most models. The percentage of students who were in their first year (mobil) was excluded from most models as well. About 10-25% 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) for both forward and backward selection. The average model size and its standard error can be found



(a) (b)

Figure 3.10: The standard error of the model-assisted estimates using variables resulting from forward and backward selection, all of the variables (full model), and none of the variables (HT estimator). The additive model is estimated with (a) Linear Splines and (b) Quadratic Splines.

in Table B.2 for both forward and backward selection methods. Forward selection had less variables than backward selection at sample size $n = 50$. At sample size $n = 200$, the average model size was about 3 for both forward and backward selection. The standard deviation 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 B.3. Models selected from both forward and backward selection had lower standard errors than using either the full model or the Horvitz-Thompson estimator for sample size $n = 100$ and $n = 200$, as seen in Figure 3.10. The bias for the backward, forward and full models ($n = 100$ and $n = 200$) had bias values nearly identical to the bias values for the Horvitz-Thompson

estimator, which we know are unbiased estimators. However, for $n = 50$, there is a negative bias for all models.

Models resulting from both the forward and backward methods reduced the standard error of thr 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% of simulations with larger sample sizes. This application demonstrates that the proposed information criterion is useful in practice.

4 Asymptotic Theory

4.1 Introduction

The inclusion of unimportant auxiliary variables in the model inflates the variance of the total estimates and parameter estimates. Variable selection methods attempt to select a parsimonious model that contains only relevant variables. Information criteria such as AIC and BIC are used to select the "best model." These information criteria penalize the model likelihood with a function of the number of parameters in the model and provide a useful metric for comparing candidate models.

An information criteria that is consistent when a complex survey design is used has not been developed when assuming an additive model. The initial derivations of the AIC and BIC do not account for a complex survey design. More recent proposals developed for linear models from Hens et al. (2006), Xu et al. (2013), and Lumley and Scott (2015) account for survey weights. Hens et al. (2006) propose an AIC for missing survey data, but do not present any asymptotic results. Xu et al. (2013) develop a BIC for sampling designs with unequal weights and consider its asymptotic properties, however the authors use a non-Bayesian derivation and do not account for the lack of independence. Lumley and Scott (2015) derive an AIC from the Kullback-Leibler divergence and a BIC with a Bayesian derivation under complex sampling designs. Both the AIC and BIC include a term that accounts for the lack of independence resulting from the sampling design. However, these aforementioned methods are developed for

linear models and do not consider the additive model. The variable selection method proposed by Wang and Wang (2011) accounts for the inclusion probabilities and assumes an additive model, but it is based on the asymptotic mean squared error of the total and a proof of its consistency is not given. In this chapter we state and prove the asymptotic consistency of our likelihood-based BIC that accounts for sampling design and assumes an additive model.

4.2 Proposed Information Criterion

We introduce the notation $M \subset \{1,...,d\}$ as the indexes of the auxiliary variables included in a candidate model. We add the subscript M to previously defined notation to clarify its association with the model containing auxiliary variables of index set M. For example, θ_M in (2.3) is the spline coefficients when only using the variables from corresponding to index set M which is of size $q_M = d_M(J_n + p)$ (see Chapter 2).

Taking the likelihood approach, we propose the following BIC for consistent variable selection for additive models discussed in Section 2.3. For a candidate model M, define

$$\text{BIC}(M) = -2n\ell(\hat{\theta}) + q_M \log(n) \tag{4.1}$$

where $\ell(\hat{\theta}) = \log L(\hat{\theta}|y_j)$ is the log-likelihood of the maximum likelihood estimate of parameter θ based on a single element. Since the expectation of one element is a mean, then the Horvitz-Thompson estimator of $\ell(\theta)$ is $\hat{\ell}(\theta) = N^{-1}\sum_{i\in S}\pi_i^{-1}\ell_i(\theta)$ [Lumley and Scott (2015)]. The N in the denominator can be substituted for $\sum_{i\in S}\pi_i^{-1}$ resulting

in the Hajek estimator, a more efficient estimator when the weights are unequal. Under the assumption of Normal model errors, the formula for the BIC can be given explicitly. This likelihood estimate is similar to the one that appears for the AIC in Hens et al. (2006).

$$\mathrm{BIC}(M) = \frac{n}{N}(\sum_{i\in S}\pi_i^{-1})\log(\mathrm{WMSE}_M) + q_M \log(n), \tag{4.2}$$

where the weighted means squared error (WMSE) for candidate model M is defined as,

$$\mathrm{WMSE}_M = \frac{\sum_{i\in S}\pi_i^{-1}(y_i - \hat{m}_M(\mathbf{x}_i))^2}{\sum_{i\in S}\pi_i^{-1}}.$$

To discuss the theoretical properties of the proposed BIC, it is necessary to introduce the following assumptions.

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

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

(A3) There exists a constant $\lambda_\pi > 0$ such that $\lambda_\pi \le \liminf_{N\to\infty}\min_{i\in U}\pi_i$.

(A4) There exists a constant $\lambda_f > 0$ such that $\lambda_f \le \liminf_{N\to\infty} n/N$.

(A5) For all $i \in U, \mathrm{E}[\varepsilon_i] = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2, \mathrm{E}[\varepsilon_i^4] = \mu_4 < \infty$, and for any $i \neq j, \varepsilon_i$ and ε_j are independent.

(A6) The support of $\mathbf{X}_i$ is $[0,1]^d$ for all $i \in U$. Furthermore, the probability density function of X is bounded away from 0 and infinity on the support, written as $0 < f_X(\mathbf{x}) < \infty, \forall \mathbf{x} \in [0,1]^d$.

(A7) 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, $k_1, ..., k_{J_n}$, is such that $\min_{j\in\{1,...,J_{n-1}\}} |k_{j+1} - k_j| / \max_{j\in\{1,...,J_{n-1}\}} |k_{j+1} - k_j| > c$ for some constant $c > 0$.

(A8) Let $\alpha_l \in \mathbb{C}^{p+1}[0,1]$, where $\mathbb{C}^{p+1}[0,1]$ is the set of $p+1$-times continuously differential functions on the $[0,1]$ interval.

Assumption (A1) is necessary to bound the variance of the estimator of σ^2. Assumptions (A2) and (A4) are common in survey literature [e.g. Robinson and Särndal (1983)] to ensure the consistency of the Horvitz Thompson estimator. In order for estimates of population quantities to be unbiased, we assume that the first order inclusion probabilities are uniformly greater than zero, as in assumption (A3). Assumptions (A5) and (A8) make general assumptions about the superpopulation model errors that are common in nonparametric estimation literature. The most important feature of (A6) is assuming a compact support. Without loss of generality, data on any bounded interval can be rescaled to unit length. Assumption (A7) ensures the number of knots increase at an appropriate rate.

Theorem 1 *Let M_0 be the indexes of the auxiliary variables in the true model. Under assumptions (A1)-(A8),*

$$\lim_{N\to\infty} \mathrm{P}(\mathrm{BIC}(M_0) \leq \mathrm{BIC}(M), \textit{ for all } M \neq M_0, M \subset \{1, ..., d\}) = 1.$$

Theorem 1 states that under regularity conditions, the proposed BIC is variable selection consistent. Therefore, the correct model will have the lowest BIC among the

candidate models with probability approaching to one as the population (and sample) size go to infinity. A proof is provided in the following section.

4.3 Proof of Consistency

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

The theoretical results will be introduced after the necessary lemmas are proven.

Lemma 2 *Under assumptions (A1)-(A8), one has*

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

Proof. Since m is the true model of the relationship between the auxiliary information, $\mathbf{x}$, to the variable of interest, y, then $\varepsilon_i = y_i - m(\mathbf{x}_i)$. Let $\xi_n = \frac{1}{N} \sum_{i \in S} \frac{\varepsilon_i^2}{\pi_i}$ and $\xi_N = \frac{1}{N} \sum_{i \in U} \varepsilon_i^2$ be two quantities calculated from the sample and population, respectively. For any fixed $\varepsilon > 0$, Chebyshev's Inequality entails that

$$\mathrm{P}(\left|\xi_n - \sigma^2\right| > \varepsilon) \leq \frac{\mathrm{E}\left|\xi_n - \sigma^2\right|^2}{\varepsilon^2}.$$

Then by adding and subtracting ξ_N and applying the Cauchy-Schwarz Inequality, one

has

$$\begin{aligned} \mathrm{E}\left|\xi_n - \sigma^2\right|^2 &= \mathrm{E}\left|\xi_n - \xi_N + \xi_N - \sigma^2\right|^2 \\ &\leq 2\left[\mathrm{E}|\xi_n - \xi_N|^2 + \mathrm{E}|\xi_N - \sigma^2|^2\right]. \end{aligned} \tag{4.3}$$

Examining the first term on the right hand side of (4.3),

$$\mathrm{E}(\xi_n - \xi_N)^2 = \mathrm{E}[\mathrm{E}[(\xi_n - \xi_N)^2|U]] = \mathrm{E}[\mathrm{Var}(\xi_n - \xi_N|U)].$$

Using Equation (16) in Dol et al. (1996) with assumptions (A1)-(A5), we can bound the design variance of the Horvitz-Thompson estimator

$$\begin{aligned} \mathrm{E}[\mathrm{Var}(\xi_n - \xi_N|U)] &\leq \mathrm{E}\left[\frac{\lambda_{\max}}{N^2}\sum_{i\in U}\frac{\varepsilon_i^4}{\pi_i^2}\right] \leq \mathrm{E}\left[\frac{\lambda_{\max}}{N^2}\sum_{i\in U}\frac{\varepsilon_i^4}{\lambda_\pi^2}\right] \\ &\leq \frac{\lambda_{\max}}{N\lambda_\pi^2}B = O(N^{-1}). \end{aligned}$$

For the second term in (4.3), one has

$$\mathrm{E}|\xi_N - \sigma^2|^2 = \mathrm{Var}(\xi_N) = N^{-2}\sum_{i\in U}\mathrm{Var}(\varepsilon_i^2) = N^{-1}(\mu_4 - \sigma^4) = O(N^{-1}).$$

Therefore, $\mathrm{E}|\xi_n - \sigma^2|^2 = O(N^{-1})$. Then for any $\varepsilon > 0$, there exists

$$C = C(\varepsilon) = \sqrt{(\mu_4 - \sigma^4 + \lambda_{\max}B\lambda_\pi^{-2})/\varepsilon}$$

such that

$$\begin{aligned} \mathrm{P}(N^{1/2}|\xi_n - \sigma^2| > C) &\leq \frac{N\mathrm{E}|\xi_n - \sigma^2|^2}{C^2} \\ &\leq \frac{N}{C^2} N^{-1}((\mu_4 - \sigma^4) + \lambda_{\max} B \lambda_\pi^{-2}) \\ &\leq \frac{\mu_4 - \sigma^4 + \lambda_{\max} B \lambda_\pi^{-2}}{C^2} \\ &= \varepsilon. \end{aligned}$$

Thus, it has been shown that

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

Lemma 3 *Under assumptions (A1)-(A8), one has*

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

Proof. Let $\mathbf{Y}_s = (Y_1, \ldots, Y_n)^T$ be the vector of responses observed in the sample. Then one can decompose $\mathbf{Y}_s$ as $\mathbf{Y}_s = \mathbf{M}_s + \mathbf{E}_s$, where $\mathbf{M}_s = (m(\mathbf{X}_1), \ldots, m(\mathbf{X}_n))^T$ and $\mathbf{E}_s = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are the mean and error vectors respectively. Define $\widehat{\mathbf{M}}_s = (\widehat{m}(\mathbf{X}_1), \ldots, \widehat{m}(\mathbf{X}_n))^T$ be the estimated mean vector based on the sample. Let $\mathbf{P}_s = \mathbf{\Gamma}_S [\mathbf{\Gamma}_S' \mathbf{\Pi}_S^{-1} \mathbf{\Gamma}_S]^{-1} \mathbf{\Gamma}_S' \mathbf{\Pi}_S^{-1}$. Then one can write

$$\widehat{\mathbf{M}}_s = \mathbf{P}_s \mathbf{Y}_s = \mathbf{P}_s \mathbf{M}_s + \mathbf{P}_s \mathbf{E}_s = \widetilde{\mathbf{M}}_s + \widetilde{\mathbf{E}}_s,$$

and

$$N^{-1}\sum_{i\in S} w_i(m(\mathbf{x}_i)-\hat{m}(\mathbf{x}_i))^2 \;\leq\; N^{-1}\left(\mathbf{M}_s-\widetilde{\mathbf{M}}_s\right)^T \mathbf{\Pi}_S^{-1}\left(\mathbf{M}_s-\widetilde{\mathbf{M}}_s\right)+N^{-1}\widetilde{\mathbf{E}}_s^T\mathbf{\Pi}_S^{-1}\widetilde{\mathbf{E}}_s$$
$$= \; I+II. \tag{4.4}$$

For I in (4.4), the approximation theorem of the polynomial spline in de Boor (2001) entails that there exists an additive spline function $g(\mathbf{x})=\sum_{l=1}^{d} g_l(x_l)$, such that

$$\max_l \sup_x |\alpha_l(x)-g_l(x)| \leq cJ_n^{p+1}$$

for some constant c that does not depends on the sample size. Let $\mathbf{G}_s=(g(\mathbf{x}_1),\ldots,g(\mathbf{x}_n))^T$. Then, by the definition of $\widetilde{\mathbf{M}}_s$, one has

$$I \leq N^{-1}(\mathbf{M}_s-\mathbf{G}_s)^T\mathbf{\Pi}_S^{-1}(\mathbf{M}_s-\mathbf{G}_s) \leq cK_n^{-2(p+1)}N^{-1}\sum_{i\in s} w_i = O_p\left(K_n^{-2(p+1)}\right). \tag{4.5}$$

For II in in (4.4), let $\left\{\varphi_j\right\}_{j=1}$ be a set of orthonormal basis of the additive polynomial spline space with respect to the empirical inner product $\langle f,g\rangle_n = N^{-1}\sum_{i\in s} w_i f(\mathbf{X}_i)g(\mathbf{X}_i)$. Then one can write $\widetilde{\mathbf{E}}_s$ as $\widetilde{\mathbf{E}}_s=\sum_{j=1}^{K_n} a_j\varphi_j$, with $a_j=\left\langle \mathbf{E}_s,\varphi_j\right\rangle_n = N^{-1}\sum_{i\in s} w_i\varphi_j(\mathbf{X}_i)\varepsilon_i$. Thus

$$II=\left\langle \widetilde{\mathbf{E}}_s,\widetilde{\mathbf{E}}_s\right\rangle_n=\sum_{j=1}^{K_n} a_j^2=\sum_{j=1}^{K_n}\left(N^{-1}\sum_{i\in s} w_i\varphi_j(\mathbf{X}_i)\varepsilon_i\right)^2$$

Note that for any $j=1,\ldots,K_n$, $N^{-1}\sum_{i\in U} w_i^2\varphi_j^2(\mathbf{X}_i)\varepsilon_i^2 \leq N^{-1}\lambda_\pi^{-2}\sum_{i\in U}\varphi_j^2(\mathbf{X}_i)\varepsilon_i^2 \leq$

$c\lambda_\pi^{-2}B$. Then by the consistency of HT estimator, one has

$$N^{-1}\sum_{i\in s} w_i\varphi_j(\mathbf{X}_i)\,\varepsilon_i = N^{-1}\sum_{i\in U}\varphi_j(\mathbf{X}_i)\,\varepsilon_i + O_p\left(\sqrt{1/N}\right).$$

Therefore,

$$II = \sum_{j=1}^{K_n}\left(N^{-1}\sum_{i\in U}\varphi_j(\mathbf{X}_i)\,\varepsilon_i\right)^2 + O_p\left(K_n/N\right),$$

in which

$$E\left(N^{-2}\sum_{j=1}^{K_n}\left(\sum_{i\in U}\varphi_j(\mathbf{X}_i)\,\varepsilon_i\right)^2 |\mathbf{X}_1,\ldots,\mathbf{X}_n\right) = N^{-2}\sum_{j=1}^{K_n}\sum_{i\in U}\varphi_j^2(\mathbf{X}_i)\,\sigma^2 = O_p\left(K_n/N\right).$$

Therefore

$$II = O_p\left(K_n/N\right). \tag{4.6}$$

The theorem follows from equations (4.4), (4.5) and (4.6). ■

Lemma 4 *Under assumptions (A1)-(A8), one has*

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

Proof. By the Cauchy-Schwarz inequality and Lemmas 1 and 2, one has

$$\begin{aligned}
&\left|N^{-1}\sum_{i\in S} w_i(y_i - m(\mathbf{x}_i))(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))\right| \\
\leq\ & \sqrt{N^{-1}\sum_{i\in S} w_i(y_i - m(\mathbf{x}_i))^2}\sqrt{N^{-1}\sum_{i\in S} w_i(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))^2} \\
=\ & O_P(\sqrt{K_n^{-2p-2} + K_n/N}). \quad ■
\end{aligned}$$

Lemma 5 *Under assumptions (A1)-(A8), one has*

$$|\hat{N}/N - 1| = O_P(N^{-1/2}),$$

where $\hat{N} = \sum_{i \in S} \pi_i^{-1}$.

Proof. Consider the expectation of $\hat{N}$,

$$\begin{aligned} E[\hat{N}] &= E[E[\hat{N}|U]] = E[E[\sum_{i \in S} \pi_i^{-1}|U]] \\ &= E[E[\sum_{i \in U} \pi_i^{-1} I_i|U]] = E[\sum_{i \in U} \pi_i^{-1} E[I_i|U]] \\ &= E[\sum_{i \in U} 1] = N. \end{aligned}$$

The variance,

$$\begin{aligned} \mathrm{Var}(\hat{N}) &= E[\mathrm{Var}(\hat{N}|U)] + \mathrm{Var}(E[\hat{N}|U]) \\ &= E[\mathrm{Var}(\hat{N}|U)] + \mathrm{Var}(N) \\ &= E[\mathrm{Var}(\hat{N}|U)] \\ &= E\left[\mathrm{Var}\left(\sum_{i \in S} \pi_i^{-1}|U\right)\right] \\ &= E\left[\sum_{i \in U} \pi_i^{-1}\pi_i(1-\pi_i)\right] \\ &\leq \sum_{i \in U} (1 - \lambda_\pi) \\ &= N(1 - \lambda_\pi). \end{aligned}$$

Therefore,

$$\mathrm{Var}(\hat{N}/N) \leq N^{-2}N(1-\lambda_\pi) = N^{-1}(1-\lambda_\pi) = O(N^{-1}).$$

It follows from Chebyshev's Inequality that for any $\varepsilon > 0$, there exists $C = C(\varepsilon) = \sqrt{(1-\lambda_\pi)/\varepsilon}$ such that,

$$\mathrm{P}(N^{1/2}|\hat{N}/N - 1| > C) \leq \frac{N\mathrm{Var}(\hat{N}/N)}{C^2} = \frac{NN^{-1}(1-\lambda_\pi)}{(1-\lambda_\pi)/\varepsilon} = \varepsilon.$$

Thus,

$$|\hat{N}/N - 1| = O_P(N^{-1/2}). \quad \blacksquare$$

Define the weighted mean squared error with respect to model M as

$$\mathrm{WMSE}_M = \frac{\sum_{i\in S} \pi_i^{-1}(y_i - m_M(\mathbf{x}_i))^2}{\sum_{i\in S} \pi_i^{-1}}.$$

Lemma 6 *Let M_0 be the true model. Under assumptions (A1)-(A8), one has*

$$|\mathrm{WMSE}_{M_0} - \sigma^2| = O_P(\sqrt{K_n^{-2p-2} + K_n/N}).$$

Proof. The numerator can be decomposed as,

$$\begin{aligned} N^{-1}\sum_{i\in S} \pi_i^{-1} e_i^2 &= N^{-1}\sum_{i\in S} \pi_i^{-1}(y_i - m(\mathbf{x}_i))^2 \\ &+ N^{-1}\sum_{i\in S} \pi_i^{-1}(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))^2 \\ &+ 2N^{-1}\sum_{i\in S} \pi_i^{-1}(y_i - m(\mathbf{x}_i))(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)) \end{aligned}$$

$$= \quad \mathbf{I} + \mathbf{II} + \mathbf{III}.$$

From the results of Lemmas 2, 3, and 4, and Slutsky's Theorem, we have that

$$|\mathbf{I} + \mathbf{II} + \mathbf{III} - \sigma^2| \leq |\mathbf{I} - \sigma^2| + |\mathbf{II}| + |\mathbf{III}| = O_P(K_n^{-2p-2} + K_n/N). \tag{4.7}$$

Combining the results from Equation 4.7 with Lemma 5 using Slutsky's Theorem yields

$$\left| \frac{N^{-1} \sum_{i \in S} \pi_i^{-1} e_i^2}{N^{-1} \sum_{i \in S} \pi_i^{-1}} - \sigma^2 \right| = O_P(K_n^{-2p-2} + K_n/N). \tag{4.8}$$

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

Let $\mathbb{H}_M$, for $M \subset \{1, ..., d\}$, be the space of all squared integrable additive functions for variables x_l, $l \in M$. Let $\mathbb{G}_M$ be the space of function 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 $q_M = 1 + r(q + J_n)$, 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 $l \in \{1, ..., d\}$, where C_l is the support of X_l. Without loss of generality, it will also be assumed that C_l is the unit interval.

Similar to Huang (1998), inner products defined on $\mathbb{H}_M$ are introduced below.

$$\begin{aligned}
\langle f,g\rangle &= E[f(\mathbf{X})g(\mathbf{X})] \\
\langle f,g\rangle_N &= \frac{1}{N}\sum_{i\in U} f(\mathbf{X}_i)g(\mathbf{X}_i) \\
\langle f,g\rangle_n &= \frac{1}{N}\sum_{i\in s} \pi_i^{-1} f(\mathbf{X}_i)g(\mathbf{X}_i)
\end{aligned}$$

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 will be used to define the orthogonal projection onto $\mathbb{G}_M$ and $\mathbb{H}_M$ as $\text{Proj}_{M,n}$ and Proj_M respectively. To reduce equation length, let $m^*_{M,n} = \text{Proj}_{M,n} m$ and $m^*_M = \text{Proj}_M m$.

Lemma 7 *Let $\mathbb{G} = \mathbb{G}_{\{1,\ldots,d\}}$ be the spline space using all available auxiliary variables, then under (A1)-(A6)*

$$\sup_{g\in\mathbb{G}} \left| \frac{\|g\|_N}{\|g\|} - 1 \right| = o_P(1).$$

This was proved in Huang (1998). A variant of the proof will be present here which closely follows the proof given in Xue and Yang (2006).

Proof. Consider the B-spline basis because it is equivalent to the truncated power basis, but with nice local properties as shown by de Boor (2001). Denote the B-spline basis for g_l as $\mathbf{b}_l = \{b_{l0},\ldots,b_{lK_n}\}$, where $K_n = J_n + p$.

Let,

$$B_{lk} = \sqrt{J_n}\left(b_{lk} - \frac{E(b_{lk})}{E(b_{l0})}\right), \quad k = 1,\ldots,K_n$$

and

$$\mathbf{B} = (B_{1,1}, ..., B_{1,K_n}, ..., B_{d,1}, ..., B_{d,K_n}).$$

By Theorem 5.4.2 in deVore and Lorentz (1993), we know that

$$E[B_{lk}] = 0, \text{ and } E|B_{lk}|^r \asymp K_n^{r/2-1}, \quad \text{for } r > 1. \tag{4.9}$$

Denote $\xi = (E_n - E)[B_{lk}(X_l)] = \frac{1}{n}\sum \xi_i$, where $\xi_i = B_{lk}^2(X_{li}) - E[B_{lk}^2(X_{li})]$.

By Minkowski's Inequality, for $r > 3$,

$$\begin{aligned} E|\xi_i|^r &\leq 2^{r-1}\left[E|B_{lk}(X_{il})|^{2r} + \{E|B_{lk}^2(X_{il})|\}^r\right] \\ &\leq 2^{r-1}\left[c_1^r K_n^r + c_2\right] \\ &\leq c^r K_n^r. \end{aligned}$$

The second moment can be bounded as,

$$E|\xi_i|^2 = E|B_{lk}^2(X_{il})|^2 - \{E|B_{lk}^2(X_{li})\}^2 \geq cK_n - c \geq cK_n.$$

Hence, there exists a constant $c > 0$ such that,

$$E|\xi_i|^r \leq c^r K_n^r \leq (cK_n^2)^{r-2} r! E|\xi_i|^2.$$

This results is called Cramer's Condition and allows the application of Theorem 1.2.2 in Bosq (1998) with Cramer's constant $c_r = cK_n^2$. Taking $h^{-1} \asymp J_n \asymp n^{\frac{1}{2p+3}} \asymp N^{\frac{1}{2p+3}}$, for all $\varepsilon > 0$,

$$P\left(\frac{1}{N}\left|\sum_{i=1}^{N}\xi_i\right| \geq \varepsilon\sqrt{\frac{\log^2(N)}{Nh}}\right) \leq 2\exp\left(-\frac{2\varepsilon^2 N^2 \frac{\log^2(N)}{Nh}}{4\sigma^2 + cK_n^2\varepsilon N\sqrt{\frac{\log^2(N)}{Nh}}}\right) \leq cN^{-N^{p/(2p+3)}}.$$

To finish the proof, examine the infinite sum of probabilities,

$$\begin{aligned}\sum_{N=1}^{\infty} P\left(\sup|\langle \mathbf{G},\mathbf{G}\rangle_N - \langle \mathbf{G},\mathbf{G}\rangle| \geq \varepsilon\sqrt{\frac{\log^2(N)}{Nh}}\right) &\leq \sum_{N=1}^{\infty}\{dK_n\}^2\left\{cN^{-N^{p/(2p+3)}}\right\} \\ &\leq \sum_{N=1}^{\infty} cN^{-2} \\ &< \infty\end{aligned}$$

The Lemma follows from applying the Borel-Cantelli Lemma to the above result. ■

A similar relationship holds for the sample norm, $\|f\|_n$.

Lemma 8 *Under assumptions (A1)-(A8),* $\sup_{g\in\mathbb{G}}\left|\frac{\|g\|_n}{\|g\|} - 1\right| = o_P(1)$.

Proof. This proof is constructed using the B-spline basis defined in the previous lemma. Consider the difference between the theoretical expectation and its corresponding Horvitz-Thompson estimator. Then by Markov's Inequality,

$$\begin{aligned}&P\left(\left|\frac{1}{N}\sum_{i\in S}B_{lk}(X_{li})\pi_i^{-1} - E[B_{lk}(X_{li})]\right| > \varepsilon\right) \\ =\ & P\left(\left|\frac{1}{N}\sum_{i=1}^{N}B_{lk}(X_{li})\pi_i^{-1}I_i(i\in S) - E[B_{lk}(X_{li})]\right| > \varepsilon\right) \\ \leq\ & \frac{1}{\varepsilon^2}E\left|\frac{1}{N}\sum_{i=1}^{N}B_{lk}(X_{li})\pi_i^{-1}I_i(i\in S) - E[B_{lk}(X_{li})]\right|^2\end{aligned}$$

The next step is to add and subtract the population based estimate of $E[B_{lk}(X_{li})]$.

$$
\begin{aligned}
& E\left|\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\pi_i^{-1} I_i(i\in S) - E[B_{lk}(X_{li})]\right|^2 \\
= \;& E\left|\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\pi_i^{-1} I_i(i\in S) - \frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li}) + \frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li}) - E[B_{lk}(X_{li})]\right|^2 \\
\leq \;& 2\left[E\left|\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\pi_i^{-1} I_i(i\in S) - \frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\right|^2\right. \\
& \left. + E\left|\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li}) - E[B_{lk}(X_{li})]\right|^2\right]
\end{aligned}
$$

By Equation 4.9, $E[B_{lk}(X_{li})] = 0$ and

$$
E\left|\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li}) - E[B_{lk}(X_{li})]\right|^2 = \frac{1}{N}E[B_{lk}(X_{li})^2] = O(N^{-2}).
$$

It remains to consider the asymptotic behavior of the difference between the population and sample based estimate. Using conditional probability and the variance bound on the Horvitz-Thompson estimator discussed previously,

$$
\begin{aligned}
& E\left|\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\pi_i^{-1} I_i(i\in S) - \frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\right|^2 \\
= \;& E\left[E\left[\left(\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\pi_i^{-1} I_i(i\in S) - \frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\right)^2 \middle| U\right]\right] \\
= \;& E\left[\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li})\pi_i^{-1} I_i(i\in S) - \frac{1}{N}\sum_{i=1}^{N} B_{lk}(X_{li}) \middle| U\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq E\left[\frac{\lambda_{\max}}{N^2}\sum_{i=1}^{N}\frac{E[B_{lk}(X_i)^2]}{\pi_i^2}\right] \\
&\leq E\left[\frac{\lambda_{\max}}{N^2\lambda_\pi}\sum_{i=1}^{N}E[B_{lk}(X_i)^2]\right] \\
&\leq E\left[\frac{1}{N}c\right] = O(N^{-1}).
\end{aligned}
$$

Therefore,

$$P(\sup|\langle \mathbf{G},\mathbf{G}\rangle_n - \langle \mathbf{G},\mathbf{G}\rangle| \geq \varepsilon) \leq K_n^2 c N^{-1} = o_P(1). \quad \blacksquare$$

Lemma 9 *Under assumptions (A1)-(A8),* $\|\hat{m}_M - m^*_{s,n}\| = O_P\left(\sqrt{K_n^{-2p-2} + K_n/N}\right)$.

The proof is obtained by applying Lemma 8 to the population version given in Huang (1998).

Lemma 10 *Under assumptions (A1)-(A8), if M underfits then* $c(M,m) = \|m^*_M - m\| > 0$.

Proof. Following the proof given in Huang and Yang (2004), if we assume that $c(M,m) = 0$ and $M \cap M_0 = M$ then $m = m^*_M \in \mathbb{H}_M$ which contradicts the M_0 being minimal. On the other hand, if we assume that $c(M,m) = 0$ and $M \cap M_0 \neq M$ then $m = m^*_M \in \mathbb{H}_M \cap \mathbb{H}_{M_0} = \mathbb{H}_{M\cap M_0}$ which also contradicts the M_0 being minimal. Therefore $c(M,m) > 0$ for the underfitting case.

Consider a set of variables x_l, $l = 1,...,d$, which contain all relevant auxiliary variables and possibly other irrelevant information. Let $M \subset (1,...,d)$ represent a model containing x_l, $l \in M$.

Proof of Theorem 1. The following proof closely follows the one given in Huang and Yang (2004).

Overfitting. Consider $M \subset \{1,...,d\}$, such that $M_0 \subset M$. Applying Lemmas 5 and 8, as well as the orthogonal projection properties of $\hat{m}_M$ and $\hat{m}_{M_0}$,

$$\text{WMSE}_M - \text{WMSE}_{M_0} = \frac{1}{N}\sum_{i\in S}\pi_i^{-1}\|\hat{m}_M - \hat{m}_{M_0}\|_n^2 = \|\hat{m}_M - \hat{m}_{M_0}\|^2(1+o_P(1)).$$

Observe that $m_M^* = m_{M_0}^* = m$ because $M_0 \subset M$. It follows from Lemma 9 and assumption (A7) that, $\|\hat{m}_M - \hat{m}_{M_0}\| \leq \|\hat{m}_M - m_M^*\| + \|\hat{m}_{M_0} - m_{M_0}^*\| = O_P(K_n^{-2p-2} + K_n/N) = o_P(1)$. Thus,

$$\begin{aligned}
\text{BIC}(M) - \text{BIC}(M_0) &= \frac{n}{N}(\sum_{i\in S}\pi_i^{-1})(\log(\text{WMSE}_M) - \log(\text{WMSE}_{M_0})) + (q_M - q_{M_0})\log(n) \\
&= n\{1+o_p(1)\}\frac{\text{WMSE}_M - \text{WMSE}_{M_0}}{\text{WMSE}_{M_0}}\{1+o_p(1)\} + (q_M - q_{M_0})\log(n) \\
&= n\frac{\text{WMSE}_M - \text{WMSE}_{M_0}}{\sigma^2\{1+o_p(1)\}}\{1+o_p(1)\} + n^{1/(2p+3)}\{1+o_p(1)\}\log(n) \\
&= n\left[\frac{\text{WMSE}_M - \text{WMSE}_{M_0}}{\sigma^2\{1+o_p(1)\}} + n^{-\frac{2p+2}{2p+3}}\log(n)\right]\{1+o_p(1)\} \\
&\geq cn\left[-O_p(K_n^{-2p-2} + K_n/N) + n^{-\frac{2p+2}{2p+3}}\log(n)\right]
\end{aligned}$$

for some constant c. For n large enough, $cn\left[-O_p(K_n^{-2p-2} + K_n/N) + n^{-\frac{2p+2}{2p+3}}\log(n)\right] > 0$. Recall that $K_n \asymp n^{1/(2p+3)}$. Therefore, $\lim_{n\to\infty}\{P(\text{BIC}_M - \text{BIC}_{M_0} > 0)\} = 1$.

Underfitting. Let $M \subset \{1,...,d\}$, such that $M_0 \cap M \neq M_0$. A lower bound for $\text{WMSE}_M - \text{WMSE}_{M_0}$ was established in Huang and Yang (2004) as,

$$\text{WMSE}_M - \text{WMSE}_{M_0} \geq c^2(M, M_0) + o_p(1).$$

The two possible cases will be considered.

Case 1: $M \cap M_0 = M$. By Lemma 5 and 8, $\text{WMSE}_M - \text{WMSE}_{M_0} = \|\hat{m}_M - \hat{m}_{M_0}\|^2(1 + o_P(1))$. Using Lemma 9 and assumption (A7) $\|\hat{m}_M - m^*_{M,n}\| = o_P(1)$ and $\|\hat{m}_{M_0} - m\| = o_P(1)$. From the triangle inequality we have that

$$\|\hat{m}_M - \hat{m}_{M_0}\| \geq \|m^*_{M,n} - m\| - \|\hat{m}_M - m^*_{M,n}\| - \|\hat{m}_{M_0} - m\| \geq \|m^*_{M,n} - m\| - o_P(1).$$

Recall that $\mathbb{G}_M \subset \mathbb{H}_M$, hence $\|m^*_{M,n} - m\| \leq \|m^*_M - m\| = c(M, m) > 0$. Therefore, $\text{WMSE}_M - \text{WMSE}_{M_0} \geq c(M, m) + o_P(1)$.

Case 2: $M \cap M_0 \neq M$. Define $M \cap M_0 = M'$. Notice $\text{WMSE}_M - \text{WMSE}_{M'} = -\frac{1}{N}\sum_{i \in S} \pi_i^{-1} \|\hat{m}_M - \hat{m}_{M'}\|_n^2$ and $\text{WMSE}_{M'} - \text{WMSE}_{M_0} = -\frac{1}{N}\sum_{i \in S} \pi_i^{-1} \|\hat{m}_{M'} - \hat{m}_{M_0}\|_n^2$ by orthogonal projection properties. By Lemma 8, $\text{WMSE}_M - \text{WMSE}_{M_0} = \|\hat{m}_M - \hat{m}_{M'}\|^2 - \|\hat{m}_{M'} - \hat{m}_{M_0}\|^2 + o_P(1)$. Lemma 9 gives us $\|\hat{m}_M - m^*_{M,n}\| = o_P(1)$, $\|\hat{m}_{M'} - m^*_{M',n}\| = o_P(1)$, and $\|\hat{m}_{M_0} - m^*_{M_0,n}\| = o_P(1)$. Applying the Triangle Inequality, $\|\hat{m}_{M'} - \hat{m}_{M_0}\| \geq \|m^*_{M'} - m^*_{M_0,n}\| - \|\hat{m}_{M'} - m^*_{M',n}\| - \|\hat{m}_{M_0} - m^*_{M_0,n}\| \geq \|m^*_{M',n} - m^*_{M_0,n}\| - o_P(1)$ and $\|\hat{m}_M - \hat{m}_{M'}\| \leq \|m^*_{M,n} - m^*_{M',n}\| + \|\hat{m}_M - m^*_{M,n}\| + \|\hat{m}_{M'} - m^*_{M',n}\| = \|m^*_{M,n} - m_{M',n}\| + o_P(1)$. Hence,

$$\text{WMSE}_M - \text{WMSE}_{M_0} \geq \|m^*_{M',n} - m^*_{M_0,n}\|^2 - \|m^*_{M,n} - m^*_{M',n}\|^2 + o_P(1).$$

Recall that the projections $m^*_{M,n}$, $m^*_{M_0,n}$, and $m^*_{M',n}$ onto $\mathbb{G}_M$, $\mathbb{G}_{M_0}$, and $\mathbb{G}_{M'}$, respec-

tively, are orthogonal. Hence $\|m^*_{M',n} - m^*_{M_0,n}\|^2 = \|m - m^*_{M',n}\|^2 - \|m - m^*_{M_0,n}\|^2$ and $\|m^*_{M,n} - m^*_{M',n}\|^2 = \|m - m^*_{M',n}\|^2 - \|m - m^*_{M,n}\|^2$. It follows that,

$$\|m^*_{M',n} - m^*_{M_0,n}\|^2 - \|m^*_{M,n} - m^*_{M',n}\|^2 = \|m - m^*_{M,n}\|^2 - \|m - m^*_{M_0,n}\|^2.$$

Due to $\mathbb{G}_M \subset \mathbb{H}_M$,

$$\|m - m^*_{M,n}\| \geq \|m - m^*_M\| = c(M,m).$$

Due to assumption (A7), $\|m - m^*_{M_0,n}\| = \rho_{M_0} = o(1)$. Hence $\text{MSE}_M - \text{MSE}_{M_0} \geq c^2(M,M_0) + o_P(1)$.

Therefore,

$$\begin{aligned} \text{BIC}(M) - \text{BIC}(M_0) &= n\log\left(1 + \frac{\text{MSE}_M - \text{MSE}_{M_0}}{\text{MSE}_{M_0}}\right) + (q_M - q_{M_0})\log(n) \\ &\geq \log\left(1 + \frac{c^2(M,M_0) + o_P(1)}{\sigma_0^2}\right) + o_P(1). \end{aligned}$$

Thus it has been shown that

$$\lim_{N\to\infty} \text{P}(\text{BIC}(M_0) \leq \text{BIC}(M), \text{ for all } M \neq M_0, M \subset \{1,...,d\}) = 1. \quad \blacksquare$$

5 Conclusion

5.1 Conclusion

Our research further develops the theory of nonparametric modeling in survey statistics. Our proposed BIC provides a consistent variable selection method for building additive models using data from complex samples. The additive model captures the unknown nonlinear relationship, while the variable selection using the BIC decreases the variance of the total estimator by removing unimportant variables.

The problem was initially investigated through simulations to determine potential solutions using the AIC and BIC to find the important variables. These simulations also provided insight into the finite sample performance of the variable selection methods. Possible equations for the information criteria were developed from earlier literature. The likelihood proposed in the AIC for linear models by Hens et al. (2006) provided a clue on how to estimate the likelihood in a design based sample, but required modifying the penalty term. The theoretical results developed by Huang and Yang (2004) were essential for proving the theorem and lemmas involving our BIC.

Proving the consistency of our proposed BIC is a challenging theoretical problem, requiring knowledge of both sampling statistics and nonparametric modeling. In addition to the increasing number of parameters when using spline approximations, we need to incorporate two sources of variation due to the superpopulation model and complex sampling design. The consistency proof of the proposed BIC provides under-

standing of its large sample properties. We also provide the theoretical derivations of the AIC and BIC, which provides understanding of the origin of their formulations and deepens our understanding of the assumptions, approximations, and possible weaknesses of these variable selection methods.

We applied our method the California Academic Perforamce Index data set. Our method produced mean estimates of the API, which demonstrated its usefulness for an applied problem. Since the population data was available for the response value, the results of the mean estimates were compared to its true value. Our variable selection method resulted in lower MSE than the Horvitz-Thompson estimate and the estimate using the full model.

Our research is very closely related to Lumley and Scott (2015), which provides an AIC and BIC for complex samples restricted to linear models. Our method assumes a nonparametric model and approximates the addive functions using splines. Lumley and Scott (2015) propose a design effect with no rigorous theoretical justification. Our theoretical development and simulation cannot justify their proposed design effect. Wang and Wang (2011) provide a method for variable selection assuming an additive model for samples from finite populations. However, our BIC is based on the likelihood rather than the asymptotic mean squared error, as assumed by Wang and Wang (2011). Furthermore, our method is consistent beyond simple random samples.

There are some weaknesses in our proposed method. First, if the assumptions of the additive model are not satisfied, such as the true model including an interaction term, the variable selection method may fail to select the correct set of variables. In practice, this problem could be alleviated by adding new variables created from the

interaction terms of existing variables. Another weakness is that our method requires more parameters than a linear model, which means it may not always be possible to estimate spline parameters for all available variables. Possible solutions could include choosing a lower degree of polynomial, decreasing the number of knots, or using prior knowledge to reduce the total number of variables considered. Lastly, our method requires a large sample size to obtain reasonable performance. However, this will not often be a problem because surveys that use complex designs usually have large sample sizes.

Future research may improve our 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 adapted to nonparametric models. Our simulations suggest that their design effect can decrease the number of correct fits. Another way to account for the design effect is to borrow ideas from longitudinal study formulations of information criteria.

Any national survey could be improved by using our method and census data on the population. For example, the National Health Interview Survey (NHIS) could use this method to investigate relationships of factors correlated with disease or the National Survey of Family Growth (NSFG) could look for links between pregnancy and socioeconomic factors. These surveys use complex designs with multiple stages or phases of sampling. Additional research could include methods to account for intraclass correlation in clusters which may improve results. A mixed model may be sufficient to estimate this effect. The second order inclusion probabilities may need to be incorpo-

rated in this model.

Another research area is examining superpopulation models that vary across strata or clusters. Further work could examine how to extend the proposed variable selection method to this framework. One challenge is to deal with the increase in the number of parameters and the number of models that must be considered.

Other further research may investigate other estimation methods. Our current proposed method is limited to polynomial spline estimates, while there are other methods that may work better for this situation. For example, polynomial splines approximate a nonparametric space using a finite parameter space, while other methods do not make such a strong reduction. Such methods may be used to create a nonparametric information criterion that is purely nonparametric at the estimation stage.

Bibliography

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716-723.

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

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-Assisted Estimation for Complex Surveys Using Penalised Splines. *Biometrika*, 92(4):831-846.

Breidt, F. J. and Opsomer, J. D. (2000). Local Polynomial Regression Estimators in Survey Sampling. *Annals of Statistics*, 28(4):1026-1053.

Breidt, F. J., Opsomer, J. D., Johnson, A. A., and Ranalli, M. G. (2007). Semiparametric Model-Assisted Estimation for Natural Resource Surveys. *Survey Methodology*, 33(1):35–44.

Brewer, K. (1963). Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics*, 5(3):93-105.

Burnham, K. P. and Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science and Business Media.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika*, 63(3):615-620.

Cassel, C.M., Särndal, C. E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley.

Chambers, R. (1996). Robust Case-Weighting for Multipurpose Establishment Surveys. *Journal of Official Statistics*, 12(1):3-32.

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. *Journal of the American Statistical Association*, 88(421):268-277.

Chen, R. and Tsay, R. S. (1993). Nonlinear Additive ARX Models. *Journal of the American Statistical Association*, 88(423):955-967.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829-836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596-610.

Cochran, W. G. (1953). *Sampling Techniques*. John Wiley and Sons.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Deming, W. (1950). *Some Theory of Sampling*. John Wiley and Sons.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American statistical Association*, 87(418):376-382.

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

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

Dorfman, A. H. and Hall, P. (1993). Estimators of the Finite Population Distribution Function Using Nonparametric Regression. *The Annals of Statistics*, 21(3):1452-1475.

Fabrizi, E. and Lahiri, P. (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.

Fan, J. (1993). Local Linear Regression Smoothers and their Minimax Efficiencies. *The Annals of Statistics*, 21(1):196-216.

Godambe, V. and Joshi, V. (1965). Admissibility and Bayes Estimation in Sampling Finite Populations. *The Annals of Mathematical Statistics*, 36(6):1707-1722.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*. John Wiley and Sons.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297-310.

Hens, N., Aerts, M., and Molenberghs, G. (2006). Model Selection for Incomplete and Design-Based Samples. *Statistics in Medicine*, 25(14):2502-2520.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663-685.

Huang, J. Z., Horowitz, J. L., and Wei, F. (2010). Variable Selection in Nonparametric Additive Models. *The Annals of Statistics*, 38(4):22–82.

Huang, J. Z. (1998). Projection Estimation in Multiple Regression with Application to Functional ANOVA Models. *The Annals of Statistics*, 26(1):242-272.

Huang, J. Z. and Yang, L. (2004). Identification of Non-Linear Additive Autoregressive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):463-477.

Isaki, C. T. and Fuller, W. A. (1982). Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77(377):89-96.

Kish, L. (1965). *Survey Sampling*. John Wiley and Sons.

Kuo, L. (1988). Classical and Prediction Approaches to Estimating Distribution Functions from Survey Data. *Proceedings of the Section on Survey Research Methods*, 280-285.

Lin, Y. and Zhang, H. H. (2006). Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models. *Annals of Statistics*, 34(5):2272-2297.

Linton, O. and Nielsen, J. P. (1995). A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration. *Biometrika*, 82(1):93-100.

Lo, L., Fernando, R., and Grossman, M. (1993). Covariance Between Relatives in Multibreed Populations: Additive Model. *Theoretical and Applied Genetics*, 87(4):423-430.

Lumley, T. (2014). Survey: Analysis of Complex Survey Samples. R Package Version 3.30.

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

Mahalanobis, P. C. (1944). On Large-Scale Sample Surveys. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 231(584):329-451.

Mammen, E., Linton, O., and Nielsen, J. (1999). The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions. *The Annals of Statistics*, 27(5):1443-1490.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558-625.

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007). Model-Assisted Estimation of Forest Resources with Generalized Additive Models. *Journal of the American Statistical Association*, 102(478):400-409.

Panse, V. G. and Sukhatme, P. V. (1954). *Statistical Methods for Agricultural Workers.* Indian Council of Agricultural Research.

Parsonnet, V., Dean, D., and Bernstein, A. (1989). A Method of Uniform Stratification of Risk for Evaluating the Results of Surgery in Acquired Adult Heart Disease. *Circulation*, 79(6):3-12.

Public Schools Accountability Act (2000). http://www.cde.ca.gov/ta/ac/pa/. Accessed: 2016-01-22.

Robinson, P. and Särndal, C. E. (1983). Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling. *Sankhya: The Indian Journal of Statistics, Series B*, 45(2):240-248.

Royall, R. M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. *Biometrika*, 57(2):377-387.

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

Salo, J., Vuokko, L., El-Sallabi, H. M., and Vainikainen, P. (2007). An Additive Model as a Physical Basis for Shadow Fading. *Vehicular Technology, IEEE Transactions on*, 56(1):13-26.

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

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461-464.

Shively, T. S., Kohn, R., and Wood, S. (1999). Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior. *Journal of the American Statistical Association*, 94(447):777-794.

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics*, 13(2):689-705.

Stone, C. J. (1994). The Use of Polynomial Splines and their Tensor Products in Multivariate Function Estimation. *The Annals of Statistics*, 22(1):118-171.

Wahba, G. (1978). Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. *Journal of the Royal Statistical Society. Series B*, 40(3):364-372.

Wang, L. and Wang, S. (2011). Nonparametric Additive Model-Assisted Estimation for Survey Data. *Journal of Multivariate Analysis*, 102(7):1126-1140.

Wang, L. and Yang, L. (2007). Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model. *The annals of statistics*, 35(6):2474-2503.

Wright, R. L. (1983). Finite Population Sampling with Multivariate Auxiliary Information. *Journal of the American Statistical Association*, 78(384):879-884.

Xu, C., Chen, J., and Mantel, H. (2013). Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data. *Survey Methodology*, 39(2):303-321.

Xue, L. (2009). Consistent Variable Selection in Additive Models. *Statistica Sinica*, 19(3):1281-1296.

Xue, L. and Yang, L. (2006). Additive Coefficient Modeling via Polynomial Spline. *Statistica Sinica*, 16(4):14–23.

Yates, F. (1946). A Review of Recent Statistical Developments in Sampling and Sampling Surveys. *Journal of the Royal Statistical Society*, 109(1):12-43.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Charles Griffin and Co. Ltd.

Yates, F. and Grundy, P. (1953). Selection Without Replacement from within Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society. Series B*, 15(2):253-261.

APPENDIX

A Simulation Tables

Table A.1: Percent of correct fitting models using variable selection in four fixed populations of size $N = 1000$. The simulation drew 100 simple random samples of size n and selected the variables for both forward and backward approaches using the proposed method. The SBLL column contain the results from Wang and Wang (2011) for comparison.

			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

Table A.2: Monte Carlo bias and standard error of the linear spline model-assisted estimators in four fixed populations of size $N = 1000$. The simulation drew 1000 simple random samples of size n and selected the variables for both forward and backward using the proposed method. The "oracle" estimates the total using the correct auxiliary variables. The Horvitz-Thompson (HT) estimator's Monte Carlo bias and variance are included for comparison.

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

Table A.3: Monte Carlo bias and standard error of quadratic spline model-assisted estimators in four fixed populations of size $N = 1000$. The simulation drew 1000 simple random samples of size n and selected the variables both forward and backward using the proposed method. The "oracle" estimates the total using the correct auxiliary variables. The Horvitz-Thompson (HT) estimator's Monte Carlo bias and variance are included for comparison.

			Forward		Backward		Oracle		HT	
Model	σ_0	n	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1	0.1	50	0.42	15.37	0.34	15.66	0.45	15.25	-7.32	176.51
		100	0.08	10.14	0.08	10.14	0.04	10.13	-3.43	125.18
		200	0.07	6.82	0.07	6.82	0.07	6.83	-0.38	85.01
	0.4	50	1.88	61.72	1.11	64.07	1.78	61.00	-5.59	182.93
		100	0.18	40.74	0.18	40.74	0.17	40.53	-3.48	131.53
		200	0.26	27.30	0.27	27.32	0.27	27.31	-0.12	87.79
2	0.1	50	0.50	21.74	0.40	22.34	0.51	21.66	8.15	264.45
		100	-0.12	14.81	-0.12	14.81	-0.13	14.80	-5.35	185.38
		200	0.12	9.47	0.12	9.47	0.11	9.47	2.05	126.64
	0.4	50	1.37	64.15	0.78	65.88	1.19	63.55	9.88	272.16
		100	-0.52	41.67	-0.49	41.64	-0.57	41.57	-5.40	190.63
		200	0.41	28.23	0.41	28.23	0.38	28.18	2.31	129.64
3	0.1	50	-1.04	36.49	0.13	16.52	0.11	16.00	10.04	156.50
		100	-0.20	10.50	-0.20	10.50	-0.21	10.50	-0.45	112.00
		200	0.05	7.08	0.05	7.08	0.04	7.07	0.32	74.53
	0.4	50	-0.81	78.46	1.38	66.56	0.79	63.00	11.77	165.60
		100	-0.65	41.21	-0.69	41.18	-0.62	41.05	-0.50	118.92
		200	0.22	27.68	0.24	27.69	0.22	27.65	0.58	78.91
4	0.1	50	6.18	80.38	-0.06	16.57	0.00	16.38	5.19	210.22
		100	-0.20	10.66	-0.20	10.66	-0.19	10.66	-5.54	143.88
		200	-0.02	6.75	-0.02	6.75	-0.02	6.75	-1.56	98.69
	0.4	50	7.45	109.44	0.29	56.97	0.41	55.12	6.65	216.92
		100	-0.52	34.73	-0.49	34.80	-0.67	34.67	-5.49	147.91
		200	0.02	22.68	0.02	22.68	0.03	22.70	-1.30	101.27

Table A.4: Percent of correct fitting models using the BIC for forward and backward approaches using the proposed information criterion based on 1000 stratified samples of size 50, 100, and 200 from a fixed population of size 1000 with noise levels 0.1 and 0.4. Four super population models were used to generate the response values from 10 possible predictors: (1) linear 2 predictors, (2) quadratic, 2 predictors (3) exponential and sin function, 3 predictors, and (4) heteroskedastic sum of sin functions, 5 predictors.

			Correct Fit Percentage			
			Linear Splines		Quadratic Splines	
Model	σ_0	n	Forward	Backward	Forward	Backward
1	0.1	50	94	93	96	92
		100	99	99	99	99
		200	99	99	100	100
	0.4	50	79	77	87	72
		100	96	95	96	96
		200	96	96	97	97
2	0.1	50	91	88	98	95
		100	98	98	99	99
		200	99	99	99	99
	0.4	50	83	80	91	80
		100	95	94	96	96
		200	98	98	97	97
3	0.1	50	88	86	91	88
		100	98	98	98	98
		200	99	99	99	99
	0.4	50	76	73	77	66
		100	95	94	95	95
		200	97	97	98	98
4	0.1	50	78	82	75	84
		100	97	97	97	97
		200	99	99	99	99
	0.4	50	72	76	61	70
		100	96	96	95	96
		200	99	99	99	99

Table A.5: Bias and standard error of the estimates from linear splines using the proposed information criterion based on 1000 stratified samples of size 50, 100, and 200 from a fixed population of size 1000 with noise levels 0.1 and 0.4. Four super population models were used to generate the response values from 10 possible predictors: (1) linear 2 predictors, (2) quadratic, 2 predictors (3) exponential and sin function, 3 predictors, and (4) heteroskedastic sum of sin functions, 5 predictors.

			Forward		Backward		Oracle		Full		HT	
Model	σ_0	n	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1	0.1	50	3.4	18.5	3.4	18.5	3.3	18.3	14.7	34.4	0.3	199.2
		100	2.1	10.8	2.1	10.8	2.1	10.8	10.2	16.5	-1.6	134.1
		200	1.4	6.9	1.4	6.9	1.4	6.9	5.8	8.3	-1.8	93.7
	0.4	50	3.7	65.0	3.8	66.4	2.9	63.2	15.1	94.7	0.0	206.6
		100	1.7	40.6	1.6	40.6	1.6	40.6	9.9	49.4	-2.4	139.6
		200	2.0	27.0	2.0	27.1	2.0	27.0	5.9	28.5	-1.6	97.0
2	0.1	50	-6.3	49.4	-6.6	50.0	-6.2	49.0	-11.3	77.0	2.9	342.0
		100	-3.3	19.9	-3.3	19.9	-3.3	19.9	-6.4	28.2	-7.7	230.2
		200	-1.8	12.7	-1.8	12.7	-1.8	12.7	-4.3	14.7	2.0	160.0
	0.4	50	-4.8	77.3	-5.3	77.8	-5.0	75.8	-10.9	117.1	2.6	348.7
		100	-3.9	44.1	-3.8	44.1	-3.8	43.7	-6.7	55.2	-8.5	235.0
		200	-1.3	28.9	-1.3	28.9	-1.2	28.8	-4.2	30.6	2.2	162.3
3	0.1	50	-3.8	42.5	-5.4	30.4	-5.6	29.8	-8.3	43.1	-0.9	187.1
		100	-2.6	14.2	-2.6	14.2	-2.7	14.2	-4.9	18.9	3.8	121.9
		200	-1.2	9.6	-1.2	9.6	-1.2	9.6	-2.7	10.3	3.2	83.8
	0.4	50	-0.8	84.7	-3.0	71.9	-3.4	67.0	-7.9	95.7	-1.1	198.3
		100	-2.9	41.3	-2.9	41.4	-3.1	41.2	-5.2	50.3	3.0	129.9
		200	-0.8	28.4	-0.8	28.4	-0.8	28.4	-2.5	29.1	3.3	89.3
4	0.1	50	0.8	121.1	-2.4	61.1	-3.0	59.7	-3.7	75.5	-11.4	263.6
		100	-3.0	31.7	-4.1	24.2	-4.2	24.2	-6.8	27.5	-2.5	177.6
		200	-1.7	14.4	-1.7	14.4	-1.7	14.4	-3.2	15.0	0.8	116.3
	0.4	50	3.9	135.6	-1.9	84.6	-2.9	81.4	-3.6	101.1	-11.6	266.7
		100	-4.4	46.2	-4.8	40.7	-4.5	40.5	-6.9	45.6	-3.3	179.8
		200	-1.5	25.6	-1.5	25.6	-1.6	25.6	-2.9	26.6	0.9	117.7

Table A.6: Bias and standard error of the estimates from quadratic splines using the proposed information criterion based on 1000 stratified samples of size 50, 100, and 200 from a fixed population of size 1000 with noise levels 0.1 and 0.4. Four super population models were used to generate the response values from 10 possible predictors: (1) linear 2 predictors, (2) quadratic, 2 predictors (3) exponential and sin function, 3 predictors, and (4) heteroskedastic sum of sin functions, 5 predictors.

			Forward		Backward		Oracle		Full		HT	
Model	σ_0	n	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1	0.1	50	4.8	19.5	4.8	19.5	4.6	19.2	17.3	48.1	0.3	199.2
		100	2.0	10.8	2.0	10.8	2.0	10.8	10.1	16.5	-1.6	134.1
		200	1.3	6.9	1.3	6.9	1.3	6.9	5.8	8.2	-1.8	93.7
	0.4	50	6.5	66.9	9.5	86.5	4.6	63.4	16.8	144.1	0.0	206.6
		100	1.5	40.8	1.4	40.8	1.3	40.6	10.1	49.4	-2.4	139.6
		200	1.9	26.9	1.9	26.9	1.8	27.0	6.2	28.3	-1.6	97.0
2	0.1	50	-3.2	25.1	-3.1	25.7	-3.3	25.0	-12.7	66.7	2.9	342.0
		100	-2.0	12.2	-2.0	12.2	-2.0	12.2	-6.1	22.4	-7.7	230.2
		200	-0.9	7.5	-0.9	7.5	-0.9	7.5	-3.8	10.5	2.0	160.0
	0.4	50	-1.7	70.2	-2.1	81.6	-2.0	69.1	-13.2	151.4	2.6	348.7
		100	-2.5	41.3	-2.5	41.3	-2.4	41.1	-6.3	52.1	-8.5	235.0
		200	-0.4	27.4	-0.4	27.4	-0.3	27.4	-3.5	29.3	2.2	162.3
3	0.1	50	1.0	61.4	-3.3	27.8	-3.4	25.9	-9.1	55.0	-0.9	187.1
		100	-2.6	14.1	-2.6	14.1	-2.6	14.0	-6.2	18.6	3.8	121.9
		200	-0.9	9.3	-0.9	9.3	-0.9	9.3	-3.5	10.0	3.2	83.8
	0.4	50	2.9	100.2	-1.5	91.1	-2.4	72.2	-9.7	149.1	-1.1	198.3
		100	-3.1	41.7	-3.1	41.7	-3.2	41.5	-6.4	50.5	3.0	129.9
		200	-0.7	28.5	-0.7	28.5	-0.6	28.4	-3.1	29.1	3.3	89.3
4	0.1	50	3.3	144.3	-7.5	55.8	-7.4	53.6	-9.2	100.8	-11.4	263.6
		100	-2.7	31.1	-3.9	23.5	-3.8	23.5	-6.1	26.8	-2.5	177.6
		200	-1.2	13.9	-1.2	13.9	-1.2	13.9	-2.7	14.8	0.8	116.3
	0.4	50	10.4	165.8	-8.9	103.5	-7.0	79.3	-9.1	153.5	-11.6	266.7
		100	-3.7	49.2	-3.8	40.3	-3.7	40.0	-6.1	44.8	-3.3	179.8
		200	-0.8	24.9	-0.8	24.9	-0.8	24.9	-2.2	26.3	0.9	117.7

Table A.7: Percent of correct fitting models using the BIC with and without the design effect for forward and backward approaches using the proposed information criterion based on 1000 stratified samples of size 50, 100, and 200 from a fixed population of size 1000 with noise levels 0.1 and 0.4. Four super population models were used to generate the response values from 10 possible predictors: (1) linear 2 predictors, (2) quadratic, 2 predictors (3) exponential and sin function, 3 predictors, and (4) heteroskedastic sum of sin functions, 5 predictors.

			Correct Fit Percentage			
			Without Design Effect		With Design Effect	
Model	σ_0	n	Forward	Backward	Forward	Backward
1	0.1	50	94	93	95	94
		100	99	99	99	99
		200	99	99	99	99
	0.4	50	79	77	80	77
		100	96	95	96	96
		200	96	96	96	96
2	0.1	50	91	88	91	89
		100	98	98	99	99
		200	99	99	99	99
	0.4	50	83	80	84	81
		100	95	94	95	94
		200	98	98	98	98
3	0.1	50	88	86	89	87
		100	98	98	98	98
		200	99	99	99	99
	0.4	50	76	73	77	75
		100	95	94	96	95
		200	97	97	97	97
4	0.1	50	78	82	79	83
		100	97	97	97	98
		200	99	99	100	100
	0.4	50	72	76	73	78
		100	96	96	97	97
		200	99	99	99	99

Table A.8: Investigation of weak signal performance of the BIC that includes the design effect. Percent of correct fitting, underfitting, and overfitting models for forward and backward approaches using the proposed information criterion with and without the design effect based on 1000 stratified samples of size 50, 100, and 200 from a fixed population of size 1000 with noise levels 0.1 and 0.4. Two super population models were used to generate the response values from 10 possible predictors: (5) 1 strong linear signal, 1 weak linear signal, (3) 1 weak exponential signal and 1 strong sin function.

				Percent of Correct Fitting Models					
				Without Design Effect			With Design Effect		
Selection	Model	σ_0	n	Correct	Underfit	Overfit	Correct	Underfit	Overfit
Forward	5	0.1	50	52	24	31	40	52	12
			100	86	6	8	74	22	4
			200	96	0	4	96	1	3
		0.4	50	4	94	26	4	95	22
			100	2	98	7	1	98	7
			200	2	98	4	2	98	4
	6	0.1	50	11	82	28	8	90	12
			100	35	61	8	28	71	2
			200	74	24	3	66	33	1
		0.4	50	3	95	28	2	96	21
			100	2	98	5	2	98	4
			200	2	98	3	2	98	3
Backward	5	0.1	50	43	22	43	37	52	17
			100	86	6	9	73	22	5
			200	96	0	4	96	1	3
		0.4	50	3	87	37	3	90	30
			100	2	97	7	1	98	7
			200	2	98	4	2	98	4
	6	0.1	50	9	74	40	7	87	16
			100	34	60	10	27	71	3
			200	74	24	3	66	33	2
		0.4	50	3	88	38	2	92	27
			100	2	97	6	2	97	5
			200	2	98	3	2	98	3

Table A.9: Percent of correct fitting models using the AIC for forward and backward approaches based on 1000 stratified samples of size 50, 100, and 200 from a fixed population of size 1000 with noise levels 0.1 and 0.4. Four super population models were used to generate the response values from 10 possible predictors: (1) linear 2 predictors, (2) quadratic, 2 predictors (3) exponential and sin function, 3 predictors, and (4) heteroskedastic sum of sin functions, 5 predictors.

			AIC Percent of Correct Fits			
			Linear		Quadratic	
Model	σ_0	n	Forward	Backward	Forward	Backward
1	0.1	50	71	68	74	72
		100	70	69	71	69
		200	67	66	70	68
	0.4	50	25	18	27	17
		100	25	21	26	22
		200	18	17	21	20
2	0.1	50	51	46	80	77
		100	56	54	71	69
		200	51	49	64	63
	0.4	50	30	24	42	32
		100	29	25	33	28
		200	25	23	27	25
3	0.1	50	48	43	62	57
		100	52	49	51	48
		200	45	44	44	42
	0.4	50	22	16	24	13
		100	25	19	26	21
		200	24	22	27	25
4	0.1	50	45	42	57	54
		100	58	55	59	57
		200	56	55	53	52
	0.4	50	33	29	36	31
		100	39	36	42	38
		200	44	42	39	38

B Application Tables

Table B.1: Percent of models including each variable in the API set from 1000 Monte Carlo simulations using stratified sampling.

Direction	p	n	cds	dnum	meals	eng.ll	mobil	col.grd	grad.sc	enroll	hsg.col
Forward	1	50	9	4	86	21	7	15	43	66	13
		100	1	1	97	13	1	10	46	84	7
		200	6	1	100	16	1	11	79	100	7
	2	50	6	3	83	16	6	13	36	53	12
		100	4	1	96	12	2	11	45	83	7
		200	7	1	100	17	2	13	80	100	7
Backward	1	50	22	17	72	34	20	37	60	75	35
		100	9	2	88	18	3	23	56	85	17
		200	7	1	98	17	1	19	79	100	13
	2	50	38	31	71	45	37	55	69	71	53
		100	9	2	87	18	4	26	58	84	21
		200	8	1	98	17	2	22	80	99	16

Table B.2: Average model size in the API set from 1000 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 B.3: Bias and standard error of the mean estimate in the API set from 1000 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