

Article

Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data

by Chen Xu, Jiahua Chen and Harold Mantel

January 2014



How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- | | |
|---|----------------|
| • Statistical Information Service | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line | 1-877-287-4369 |

Depository Services Program

- | | |
|------------------|----------------|
| • Inquiries line | 1-800-635-7943 |
| • Fax line | 1-800-565-7757 |

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for
Statistics Canada

© Minister of Industry, 2014.

All rights reserved. Use of this publication is governed by the
Statistics Canada Open Licence Agreement ([http://www.
statcan.gc.ca/reference/licence-eng.html](http://www.statcan.gc.ca/reference/licence-eng.html)).

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

- | | |
|----------------|--|
| . | not available for any reference period |
| .. | not available for a specific reference period |
| ... | not applicable |
| 0 | true zero or a value rounded to zero |
| 0 ^s | value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded |
| P | preliminary |
| r | revised |
| X | suppressed to meet the confidentiality requirements of the <i>Statistics Act</i> |
| E | use with caution |
| F | too unreliable to be published |
| * | significantly different from reference category ($p < 0.05$) |

Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data

Chen Xu, Jiahua Chen and Harold Mantel¹

Abstract

Regression models are routinely used in the analysis of survey data, where one common issue of interest is to identify influential factors that are associated with certain behavioral, social, or economic indices within a target population. When data are collected through complex surveys, the properties of classical variable selection approaches developed in i.i.d. non-survey settings need to be re-examined. In this paper, we derive a pseudo-likelihood-based BIC criterion for variable selection in the analysis of survey data and suggest a sample-based penalized likelihood approach for its implementation. The sampling weights are appropriately assigned to correct the biased selection result caused by the distortion between the sample and the target population. Under a joint randomization framework, we establish the consistency of the proposed selection procedure. The finite-sample performance of the approach is assessed through analysis and computer simulations based on data from the hypertension component of the 2009 Survey on Living with Chronic Diseases in Canada.

Key Words: Variable selection; Sampling weights; Model-design-based inference; BIC; Penalized likelihood; Selection consistency.

1 Introduction

In many areas of scientific research, one common interest is to identify the influential factors associated with certain behavioral, social, or economic indices within a target population. For example, sociologists would like to identify important factors that affect the unemployment rate in a specific region, and epidemiologists are interested in finding risk behavior for diseases. In such studies, researchers often start with a survey of the target population (*e.g.*, Rahiala and Teräsivirta 1993; Korn and Graubard 1999; Wolfson 2004). A representative sample is then selected and measurements of the variables of interest for the sampled units are collected. A regression model is routinely employed to summarize the information contained in the data. It explains variations in the response variable through a simple function of explanatory variables (covariates). When they lack prior knowledge, researchers may collect information on many potential explanatory variables. The goal of identifying influential factors can be achieved through a variable selection procedure.

Variable selection is fundamental in statistical modeling. In non-survey settings, classical selection criteria have been developed to assess and select candidate variables. Examples include Mallows's C_p statistic (Mallows 1973), the (generalized) cross-validation (CV/GCV; Stone 1974; Craven and Wahba 1979), the Akaike information criterion (AIC; Akaike 1973) and the Bayesian information criterion (BIC; Schwarz 1978). All these criteria are very useful and can provide meaningful inferences in practice.

Despite the abundance of the literature on variable selection, it has received little attention in the context of survey sampling. When variable selection methods are applied to survey data, many potential complications arise. We focus on issues related to special features of surveys. First, data collected through

1. Chen Xu and Jiahua Chen, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4. E-mail: chen.xu@stat.ubc.ca and jhchen.stat.ubc.ca; Harold Mantel, Statistical Research and Innovation Division, Statistics Canada, Ottawa, ON, Canada, K1A 0T6. E-mail: Harold.Mantel@statcan.gc.ca.

survey sampling are usually obtained from a finite population without replacement, and hence they have an intrinsic dependence structure. Second, in complex survey designs, the inclusion probabilities of sampling units often vary over the target population. Consequently, the correlation between the response and the covariates reflected in the sample can be distorted from the population. This is potentially the case when some parts of the population are sampled more intensively than the others. Ignoring survey designs in the selecting process may result in biased selection results for the target population.

In the literature, sampling weights are often utilized in estimating parameters in regression models based on survey data. The weighted estimates of regression coefficients are helpful to avoid the biased inference from informative sampling (Pfeffermann 1993; Fuller 2009, Section 6.3; Skinner 2012). Although model estimation and selection serve for their own purposes, they often have coherent linkage in a modeling process. It is natural to conjecture that using sampling weights is beneficial for the variable selection.

In this spirit, we investigate the use of pseudo-likelihood to take account of the sampling weights, and derive a pseudo-likelihood-based BIC criterion for variable selection of survey data. A penalized pseudo-likelihood-based procedure (PPL) is further proposed for numerical implementation of the proposed criterion. Under a joint randomization framework, we prove that the new procedure consistently identifies the influential variables. The weighted selection method is assessed through simulation studies and using data from the 2009 Survey on Living with Chronic Diseases in Canada.

The paper is organized as follows. In Section 2, we introduce the joint randomization mechanism and the super-population model. In Section 3, we derive the pseudo-likelihood-based BIC for the analysis of survey data and propose its implementation via the PPL procedure. In Section 4, we investigate the asymptotic behavior of the proposed BIC procedure. We use numerical studies in Section 5 to further assess the performance of our approach and provide concluding remarks in Section 6. We provide the proofs of theorems in a separate technical supplement: Xu and Chen (2012), where the derivation of proposed BIC can also be found.

2 Joint inference and super-population

The random behavior of an inference procedure is mostly inherited from the randomness in the data. In the context of surveys, the set of sampled units is random because of the probabilistic sampling design. At the same time, the value of each sampling unit may be regarded as a random outcome from some conceptual infinite super-population (Royall 1976).

In a design-based analysis, the finite population is regarded as nonrandom and all measurements of sampling units are constants. The parameters of interest are finite population quantities such as the population total or the population median. The statistical inference is evaluated based on the randomness from the probability design.

One may also regard the design-induced randomness as an artifact. The measurements of sampled units are independent realizations of a random variable from a probability model for the postulated super-population. The parameters of interest are related to the assumed model and model-based inferences are evaluated solely based on the randomization introduced from the model.

A third approach is called model-design-based inference; it incorporates the randomization from both design and model. In such a joint randomization mechanism, the finite population is regarded as a random

sample from a super-population. The survey sample is considered as a second-phase sampling from the super-population. The parameters of interest can be either model or finite-population parameters. In this mechanism, inferences on the finite-population parameters are motivated from the super-population model. Model-design-based inference can be more efficient than pure design-based approaches when the finite population is well described by the super-population model. Compared with pure model-based approaches, it protects against model violation and is therefore more robust in general (see, *e.g.*, Binder and Roberts 2003; Kalton 1983).

We study the variable selection problem under the joint randomization mechanism. Let $\mathcal{D} = \{1, \dots, N\}$ be a finite population consisting of N sampled units. The measurements on the i^{th} unit are denoted (y_i, \mathbf{x}_i) , where y_i is the response of interest and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ is a p -dimensional explanatory vector (covariate vector). These are regarded as independent realizations of (Y, \mathbf{X}) from a super-population. We postulate a generalized linear model (GLM) on the super-population as follows. Conditioning on \mathbf{X} , the distribution of Y belongs to a natural exponential family, the density of which takes the form

$$f(y; \theta) = c(y) \exp\{\theta y - b(\theta)\}. \quad (2.1)$$

θ is known as the natural parameter of $f(y; \theta)$ such that $b'(\theta) = E[Y|X] \equiv \mu$ and $b''(\theta) = \text{Var}[Y|X] \equiv \sigma^2$, and $c(y)$ is a non-negative base measure. The influence of the explanatory variable \mathbf{X} on Y is expressed through $g(\mu) = \mathbf{X}^T \boldsymbol{\beta}$ for some assumed linkage function $g(\cdot)$, where the vector $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\}^T$ is the p -dimensional regression coefficient. If $g(\cdot)$ is the canonical link, *i.e.*, $g(\mu) = \theta$, then we have $\theta = \mathbf{X}^T \boldsymbol{\beta}$. For simplicity, we focus on the canonical link in this paper.

Based on this model, the effect of the explanatory variable is characterized through the size of the corresponding regression coefficient. In applications, a complex model with many variables often leads to over-fitting and a poor interpretive value. Hence, it is desirable to fit the data with a parsimonious model in which many regression coefficients are estimated to be zero. Explanatory variables with nonzero coefficients are then considered to be influential on the response. To this end, we assume that $\boldsymbol{\beta}$ is ideally sparse, and address the variable selection problem through identifying a sparse model formed by the covariates with nonzero coefficients.

3 Pseudo-likelihood-based selection with BIC

3.1 BIC in surveys

With the model settings described in Section 2, it is clear that, if the measurement (y_i, \mathbf{x}_i) is observed for every unit in population \mathcal{D} , the randomness in the data introduced by the probability sampling design is completely gone. In this situation, the selection of the influential variables is based on the entire population and the classical selection criteria developed in non-survey settings (purely model-based) remain valid for model-design-based inference. In particular, let $s \subseteq \{1, \dots, p\}$ be an arbitrary set of

$\tau(s)$ covariates, which corresponds to a candidate model in form of (2.1). The ‘‘census-based’’ BIC (Schwarz 1978) selects the model (covariates) that minimizes

$$\text{BIC}_N(s) = -2l_N(\tilde{\boldsymbol{\beta}}_s) + \tau(s)\log N, \quad (3.1)$$

where $l_N(\boldsymbol{\beta}) = \sum_{i=1}^N \log f(y_i; \mathbf{x}_i; \boldsymbol{\beta})$ is the census log-likelihood function and $\tilde{\boldsymbol{\beta}}_s$ is the maximizer of $l_N(\boldsymbol{\beta})$ based on s . It can be seen that the BIC (3.1) is a decreasing function of the maximized log-likelihood and an increasing function of the number of variables included in the model. Hence, a lower BIC implies either a simpler model (fewer explanatory variables), a better fit (higher maximized likelihood), or both. A model with balanced complexity and goodness of fit is preferred.

We note that the census BIC (3.1) is conceptual, because observing (y_i, \mathbf{x}_i) for all units in \mathcal{D} is usually not feasible in applications. Instead, a representative sample $d = \{i_1, \dots, i_n\} \subset \{1, \dots, N\}$ with n units is often drawn from \mathcal{D} and the measurements are observed based on the sampled units. Due to the intrinsic dependence structure among the sampled units, a full likelihood on d is prohibitive to compute in general. Alternatively, for the model-design-based inference, a pseudo-log-likelihood function is frequently used, which takes the form

$$l_n(\boldsymbol{\beta}) = \sum_{i \in d} w_i \log f(y_i; \boldsymbol{\beta}) \quad (3.2)$$

with $w_i = k/P(i \in d)$ denoting the survey weight for the i^{th} unit. The scaling parameter k in w_i does not have analytical impacts on the pseudo-likelihood-based inference. For the simplicity of presentation, we choose $k = n/N$ such that $n^{-1}l_n(\boldsymbol{\beta})$ is design-unbiased to $N^{-1}l_N(\boldsymbol{\beta})$. Maximizing $l_n(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ leads to a maximum pseudo-likelihood estimator (MPLE) $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, *i.e.*,

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} l_n(\boldsymbol{\beta}).$$

Under the appropriate sampling designs, $\hat{\boldsymbol{\beta}}$ is often $n^{-1/2}$ consistent for $\boldsymbol{\beta}$ under the joint randomization framework. The idea of using pseudo-likelihood for inference on model parameters has been widely adopted in the literature (see, *e.g.*, Binder 1983; Godambe and Thompson 1986; Molina and Skinner 1992).

In this paper, we aim to develop an analogue of BIC criterion based on the pseudo-likelihood. Following the super-population formulation described in Section 2, let $\boldsymbol{\beta}_s$ be the $\tau(s)$ -dimensional coefficient of model s and let ν_s be the prior density of $\boldsymbol{\beta}_s$. Then a pseudo-marginal density function of the data is given by

$$P_n(\mathbf{y}|s) = \int L_n(\mathbf{y}; \boldsymbol{\beta}_s) \nu_s(\boldsymbol{\beta}_s) d\boldsymbol{\beta}_s$$

with $L_n(\mathbf{y}; \boldsymbol{\beta}_s) = \exp\{l_n(\mathbf{y}; \boldsymbol{\beta}_s)\}$. Consequently, we may regard the following expression as the pseudo-posterior probability of the model s :

$$P_n(s|\mathbf{y}) = \frac{P_n(\mathbf{y}|s)P(s)}{\sum_{s \in \mathcal{S}} P(s)P_n(\mathbf{y}|s)}, \quad (3.3)$$

where S denotes the collection of all candidate models. In the spirit of Bayesian analysis, the model with the highest $P_n(s|\mathbf{y})$ is then considered to be the one that receives the most support from the data. Since $\sum_{s \in S} P(s) P_n(\mathbf{y}|s)$ does not depend on any specific model, the highest $P_n(s|\mathbf{y})$ is achieved by the model that maximizes the corresponding $P_n(\mathbf{y}|s)P(s)$. When the uniform prior $P(s) = \zeta$ is used and the weight scaling is chosen as $k = n/N$, we obtain a Laplace approximation under some regularity conditions (see Xu and Chen 2012):

$$-2 \log \{P_n(\mathbf{y}|s)\} = -2l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s) \log n + O_p(1).$$

Accordingly, we choose the model s that minimizes

$$\text{BIC}_n(s) = -2l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s) \log n. \quad (3.4)$$

Compared with the census BIC (3.1), the first term in BIC (3.4) is the maximum survey-weighted pseudo-likelihood, which is potentially helpful to avoid sampling errors that might lead to biased inferences for the target population. We refer to (3.4) as a pseudo-likelihood-based version of BIC in the context of surveys. In the joint randomization framework, we establish the selection consistency of using BIC (3.4) through a PPL-based implementation procedure, as will be seen in Section 4.

3.2 Implementing BIC via penalized pseudo-likelihood

In applications, a straightforward way to implement BIC is best-subset selection, where BIC is evaluated and compared for each candidate model. However, this procedure can be computationally impractical when the number of covariates is large. Alternatively, penalized likelihood methods have recently been used as computationally efficient procedures for implementing a selection criterion. These methods exclude variables from the model by estimating their coefficients to be zero, and shrink the other coefficients accordingly. By varying the penalty on the likelihood, we can obtain a series of models with differing sparsity. To avoid an exhaustive search of the entire model space, the selection criterion is used to pick an optimal one among these sparse models. The effectiveness of this implementation strategy has been illustrated in the non-survey context for BIC (Wang, Li and Tsai 2007; Liu, Wang and Liang 2011) and GCV (Fan and Li 2001; Xie, Pan and Shen 2008) among others.

Sharing the same spirit, we proposed a penalized pseudo-likelihood (PPL) procedure for the implementation of BIC (3.4) for survey data. Specifically, following pseudo-likelihood (3.2) with $k = n/N$, we define the survey-weighted penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$ that maximizes the penalized pseudo-likelihood function

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n \sum_{j=1}^p \phi_\lambda(|\beta_j|), \quad (3.5)$$

where $\phi_\lambda(\cdot)$ is a penalty function indexed by a tuning parameter λ controlling the size of the penalty. With an appropriate choice of $\phi_\lambda(\cdot)$, $\hat{\boldsymbol{\beta}}_\lambda$ contains zero estimates for some coefficients and thus automatically produces a sparse model. The desirable sparsity of $\hat{\boldsymbol{\beta}}_\lambda$ typically requires the singularity of the corresponding $\phi_\lambda(\cdot)$ at the origin. Some popular choices of $\phi_\lambda(\cdot)$ include the L_γ penalty (Frank and

Friedman 1993; Tibshirani 1996), *i.e.*, $\phi_\lambda(|\beta|) = \lambda |\beta|^\gamma$ with $\gamma \in (0, 1]$, and the SCAD penalty (Fan and Li 2001), which is defined by the following derivative:

$$\phi'_\lambda(|\beta|) = \lambda \left\{ I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\} \quad (3.6)$$

with $a = 3.7$ being a common choice.

With different values of λ for a properly specified $\phi_\lambda(\cdot)$, $\hat{\beta}_\lambda$ leads to models of differing sparsity. These sparse models (with respect to λ) naturally form a collection of candidate models. BIC (3.4) can then be used to select an optimal model within this collection. To be more specific, let Ω be the range of λ and let s_λ denote the model produced by $\hat{\beta}_\lambda$. We treat $S_\Omega = \{s_\lambda : \lambda \in \Omega\}$ as the collection of candidate models under consideration, and select the model $s^* \in S_\Omega$ such that $\text{BIC}_n(s^*) = \min_{\lambda \in \Omega} \text{BIC}(s_\lambda)$. We refer to this selection procedure as the penalized pseudo-likelihood-based BIC method (PPL-BIC). Compared with traditional best-subset selection, the PPL-BIC procedure focuses on the models that are produced by the survey-weighted penalized estimators, and therefore it can be much less computationally expensive.

4 Consistency of PPL-BIC

We now investigate the asymptotic behavior of the PPL-BIC procedure under the joint randomization framework. Suppose there is a sequence of finite populations, say \mathcal{D}_r with $r \rightarrow \infty$. Each \mathcal{D}_r is an independent and identically distributed (i.i.d.) sample of size N_r from a super-population modeled by (2.1) with random variable $(Y, \mathbf{X} = \{X_1, \dots, X_p\})$. Within each \mathcal{D}_r , a sample d_r of size n_r is drawn according to some sampling scheme. We assume that both N_r and n_r increase to infinity as $r \rightarrow \infty$, with the sampling fraction n_r/N_r bounded by some constant $C < 1$. For simplicity of notation, we will drop the index r in the following discussion.

Without loss of generality, we assume that the first q coefficients are nonzero and denote the true value of β by $\beta_0 = \{\beta_{01}, \beta_{02}\}$ with $\beta_{02} = 0$. Also, we use s_0 to denote the true model $\{1, \dots, q\}$ to be identified. We establish the selection consistency of PPL-BIC in two steps. In the first step we show that, for appropriate choices of $\phi_\lambda(\cdot)$, the PPL can consistently identify the true s_0 so that $s_0 \in S_\Omega$ with probability tending to 1. In the second step, we verify that BIC (3.4) consistently selects s_0 over S_Ω .

For the asymptotic analysis, we define $\varphi_\lambda = \max \left\{ \phi'_\lambda(|\beta_{0j}|) \text{ for } j \in s_0 \right\}$ and associate λ with n to make φ_λ a sequence. Under the joint randomization framework, we show the claim of step 1 as the following theorem.

Theorem 1 Under regularity conditions on model (2.1) and other requirements specified in the online supplement, if $\varphi_\lambda \rightarrow 0$ as $n \rightarrow \infty$, then there exists a local maximizer $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda 1}, \hat{\beta}_{\lambda 2})$ of the penalized pseudo-likelihood function (3.5) such that

$$\|\hat{\beta}_\lambda - \beta_0\| = O_p(n^{-1/2} + \varphi_\lambda) \quad \text{and} \quad P\{\hat{\beta}_{\lambda 2} = 0\} \rightarrow 1$$

with $\|\cdot\|$ denoting the Euclidean norm.

The consistency result in Theorem 1 holds for popular nonconvex penalty functions. For example, for the L_γ penalty with $\gamma \in (0,1)$, consistency holds if $\lambda \rightarrow 0$; for the SCAD penalty, consistency holds if $\lambda \rightarrow 0$ and $\sqrt{n}\lambda \rightarrow \infty$. It also implies that with probability tending to 1, the true model s_0 is included in S_Ω , which serves as a prerequisite for the selection consistency of BIC over S_Ω .

We now establish the consistency of using BIC on S_Ω with a specified $\phi_\lambda(\cdot)$ that satisfies Theorem 1. Following the notation used in Section 3.2, let s_λ be the model corresponding to a PPL estimator $\hat{\beta}_\lambda$, and let Ω be the range of λ under consideration. We define two collections of candidate models as follows:

- Over-fitted models: $S_+ = \{s : s_0 \subset s, s \neq s_0\}$;
- Under-fitted models: $S_- = \{s : s_0 \not\subset s\}$.

Notation $\not\subset$ denotes there is at least one different element between two sets, so that S_- is the collection of candidate models which does not include all variables in the true model. Then, Ω can be partitioned accordingly into

$$\Omega_+ = \{\lambda : s_\lambda \in S_+\}, \quad \Omega_- = \{\lambda : s_\lambda \in S_-\}, \quad \Omega_0 = \{\lambda : s_\lambda = s_0\}. \quad (4.1)$$

By Theorem 1, we have shown that $P(\Omega_0 \neq \emptyset) \rightarrow 1$. Therefore, the selection consistency of BIC over S_Ω is achieved if BIC is able to identify s_0 from any model s_λ with $\lambda \in \Omega_+ \cup \Omega_-$. We use the following theorem to establish this consistency result.

Theorem 2 Under the same conditions as in Theorem 1,

$$P\left\{\min_{\lambda \in \Omega_+ \cup \Omega_-} \text{BIC}_n(s_\lambda) \leq \text{BIC}_n(s_0)\right\} \rightarrow 0,$$

where Ω_+ and Ω_- are defined in (4.1).

5 Numerical studies

To evaluate the finite sample performance of PPL-BIC, extensive numerical studies have been conducted using data from the Survey on Living with Chronic Diseases in Canada (SLCDC; Statistics Canada 2009). In particular, we compare the proposed procedure with classic non-survey methods based on regression models postulated between SLCDC variables and hypothetical (simulated) responses. We tentatively reveal some insights for using pseudo-likelihood-based selection under two simulation scenarios. In the first scenario, populations are generated from presumed models and samples are obtained by designs that potentially create spurious correlations among SLCDC variables. In the second scenario, populations are not accurately generated from presumed models and samples are obtained by a design related to both response and candidate covariates. Also, we report the analysis of the original SLCDC 2009 data as an example for using PPL-BIC in real applications.

5.1 SLCDC data

SLCDC is a cross-sectional study sponsored by the Public Health Agency of Canada that collects information related to the experiences of Canadians with chronic health conditions. One of the main

objectives of SLCDC is to identify health behavior that influences disease outcomes, so that the government can better plan and provide health services for people with chronic diseases.

SLCDC takes place every two years, with two chronic diseases covered in each survey cycle. The 2009 survey focused on arthritis and hypertension. We restrict our attention to hypertension. The target population for the hypertension survey is Canadians aged twenty years or older from the ten provinces who have been diagnosed with hypertension and who live in private dwellings. To facilitate the survey process, the sampling units of SLCDC 2009 are people with hypertension who completed the 2008 Canadian community health survey (CCHS). For the purpose of SLCDC, the population is first stratified according to the CCHS respondents based on sex and four age groups: 20-44, 45-64, 65-75, and 75+. Therefore, the finite population formed by the CCHS respondents was divided into 8 categories, age (4 levels) by sex (2 levels). A stratified sampling plan is used for SLCDC with proportional sample size allocation. An overall sample of 9,005 was selected from the 17,437 CCHS respondents, and 6,142 respondents completed the SLCDC survey.

We identified 40 variables relevant to hypertension based on the original SLCDC data, among which 7 variables have complete information on all 6,142 respondents. The remaining 33 variables have some amount of missing values due to the non-responses in the original questionnaire (see Table 5.5 in Appendix for the list of variables and corresponding non-response rates). There was no obvious systematic reason for the item non-response. The variable with most severe missingness is INCDRPR (household income) with a 9.6% non-response rate, while the amount of missing data is relatively minor for the remaining variables. To facilitate the analysis, we used simple imputation methods for the missing data as follows. For a categorical variable, we imputed the non-response value by a random value from the response set; for a continuous variable, we imputed the non-response value by the mean value of the responses. Two exceptions for above imputation are variables BMHX_02 and CNHX_05. The former one acts as the response variable of the regression model in the later data analysis, while the later one has natural restrictions on the range of its value. Instead, we removed the 274 observations with missing values in these two variables, which results in the basic working data with 5,868 observations. The imputation/removal procedure does not have any effect on evaluating the BIC procedure based on simulated population. It could bias the analysis of the real data. Yet given the low rate of missingness, and plausibility of missing at random in the specific case, the conclusion is unlikely to be severely affected.

Since the SLCDC is a follow-up to the CCHS, the sampling weights for SLCDC were initially obtained from the weights of the CCHS data. The weights were then adjusted to ensure that the SLCDC respondents represent the target population. Consequently, the adjusted weights show considerable variation between sampled units. After scaling by $k = n/N \approx 10^{-3}$, the adjusted weights vary between 0.01 to 33.62 with an inter-quartile range of 0.76.

5.2 Scenario 1: Spurious correlation

As mentioned, in complex survey designs, the correlation structure between variables reflected in the sample can be distorted from the population. In the first simulation scenario, we assess the proposed BIC method when data are collected through designs that potentially create spurious correlations between candidate covariates. Specifically, we treat the 40 identified variables as candidate covariates for some hypothetical response Y , and index them as X_1 to X_{40} for simplicity. We consider both continuous and binary responses in our simulations. For the continuous cases, we generate the values of Y according to

- Model 1 : $Y = 0.7X_6 + 0.7X_{10} + 0.6X_{18} - 0.6X_{22} + \varepsilon$,
- Model 2 : $Y = 0.7X_6 + 0.6X_{10} + 0.6X_{18} - 0.5X_{22} + 0.3X_{30} - 0.3X_{34} + \varepsilon$,

with $\varepsilon \sim N(0,1)$. For the binary cases where $Y \in \{0,1\}$, we generate the values of Y according to the logistic models

- Model 3 : $\text{logit}\left(\Pr\{Y = 1|\mathbf{X}\}\right) = 0.7X_7 - 0.6X_8 + 0.5X_{26}$,
- Model 4 : $\text{logit}\left(\Pr\{Y = 1|\mathbf{X}\}\right) = 0.8X_7 - 0.7X_8 + 0.6X_{26} - 0.5X_{28} + 0.4X_{36}$.

The specified models include one of the strata identifiers in SLCDC (*i.e.*, X_6 or X_7) with a nested structure for each modeling context.

The finite population used in the simulation was created as follows. The basic working data of 5,868 respondents was duplicated 10 times proportional to the rounded integer values of SLCDC weights, which results in a pseudo-finite population of size 55,950 with complete information on X_1, \dots, X_{40} . The values of response Y were then generated based on Models 1-4 respectively. We consider the variable selection problem to be the identification of the postulated model that generates the values of Y .

We investigate the performance of proposed procedure under two stratified sampling plans. Specifically, we create 4 strata based on variables X_6 (age, 55-/55+) and X_7 (sex, Male/Female), which leads to the group (Female, 55-) of size 7,120, group (Female, 55+) of size 19,199, group (Male, 55-) of size 6,187, and group (Male, 55+) of size 23,458. In the first plan, a simple random sampling without replacement (SRSWOR) with equally allocated sample size is drawn from each stratum. The inference is made based on the four SRSWORs pooled together. In the second plan, we further construct three subgroups within each stratum based on the sum of two binary covariates of the postulated models. In particular, the subgroups are built according to $X_{18} + X_{22}$ for data generated by Models 1-2, while the subgroups are similarly construct based on $X_8 + X_{26}$ for data from Models 3-4. We then make inference based on SRSWORs drawn from each sub-group of the four strata. The overall sample size is equally allocated at the stratum level with a 2:1:2 proportion for the three subgroups within a same stratum. A simple Monte Carlo computation reveals that the sample correlation between X_{18} and X_{22} (for data from Models 1-2) can be as high as 0.5, whereas their population-based correlation is merely around 0.02. Similar phenomenon is also observed between X_8 and X_{26} (for data from Models 3-4). We therefore expect variable selection under the second sampling plan to be more challenging due to this systematic inflation. In the simulations, we set the overall sample size $n = 500$ for Models 1-2 and $n = 1,500$ for Models 3-4. A summary of influential variables to the response and the design variables affecting the sampling probabilities can be found in Appendix (Table A.2).

The PPL-BIC selection procedure was carried out on probability samples obtained from the finite population. In particular, we scaled the survey weights as mentioned in (3.2) and chose the SCAD penalty for the penalized pseudo-likelihood function (3.5). The corresponding maximizer of (3.5) was solved by using the thresholding-based iterative algorithm (She 2011). For comparison purpose, the ideas of AIC and GCV are also used as alternatives for the proposed BIC (3.4). Based on the discussion in Section 3, we define the pseudo-likelihood-based AIC and GCV as

$$\text{AIC}_n(s) = -2l_n(\hat{\beta}_s) + 2\tau(s),$$

$$\text{GCV}_n(s) = -\frac{1}{n} \frac{l_n(\hat{\beta}_s)}{(1 - \tau(s)/n)^2},$$

which are similarly implemented though the PPL-based procedure. Moreover, for each setup, we repeat the selection procedure with all survey weights ignored (being set as unity). The unweighted selection results are corresponding to pure model-based inferences as discussed in Section 2. In particular, the pseudo-likelihood-based BIC reduces to the classic BIC (3.1) used for non-survey situations.

In Tables 5.1-5.2, we summarize the simulation results based on 1,000 repetitions in terms of the positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR), and averaged model size (AMS). Specifically, let s_0 be the true model that generates the finite population and s'_j be the selected model based on the j^{th} sample, $j = 1, \dots, 1,000$. The PSR, FDR, CSR and AMS are estimated as

$$\text{PSR} = \frac{\sum_{j=1}^{1,000} \tau(s_0 \cap s'_j)}{1,000 \tau(s_0)}, \text{FDR} = \frac{\sum_{j=1}^{1,000} \tau(s'_j/s_0)}{1,000 \tau(s'_j)},$$

$$\text{CSR} = \frac{\sum_{j=1}^{1,000} I(s'_j = s_0)}{1,000}, \text{AMS} = \frac{\sum_{j=1}^{1,000} \tau(s'_j)}{1,000},$$

where $\tau(s)$ denotes the size of model s and $I(\cdot)$ is the indicator function. In addition, we assess the predictive accuracy of the selected model as follows. For each setup, a test sample of size 200 is generated by SRSWOR from the same finite population as that for the training sample. For Models 1-2, we use the averaged residual sum of squares (RSS) on the test data as a measurement of the predictive ability of the selected model. For Models 3-4, we compute both positive and negative prediction rates. To be specific, let π^* be a specified benchmark and $\hat{\pi}_i$ be the estimated success probability of the i^{th} test sample, $i = 1, \dots, 200$. We then predict the i^{th} response y_i by $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi^*$ and $\hat{y}_i = 0$ otherwise. The correct prediction rates are estimated by

$$\text{PPR} = \frac{\sum_{i \in \{i: \hat{y}_i = 1\}} I(\hat{y}_i = 1)}{200}, \text{NPR} = \frac{\sum_{i \in \{i: \hat{y}_i = 0\}} I(\hat{y}_i = 0)}{200}.$$

$$\sum_{i=1}^{200} I(y_i = 1), \sum_{i=1}^{200} I(y_i = 0)$$

The final PPR and NPR are averaged based on 1,000 replications. Note that PPR and NPR here are similar to sensitivity and specificity in the clinical studies, which indicate the ability of a 0-1 prediction approach in terms of correct positive and negative predictions. In general, a larger π^* leads to high NPR but low PPR. The value of π^* should be cautiously specified in applications. In our simulation studies, we fix $\pi^* = 0.5$ for simplicity.

The results are encouraging for the proposed BIC method. From Tables 5.1-5.2, we observe that models selected by AIC have both high PSR and FDR, which indicates an excessive inclusion of the irrelevant variables. In comparison, the BIC significantly reduces the FDR of selected models with a slight sacrifice on PSR, and selects the model with sizes closer to the truth. Although the GCV behaves similarly

to BIC in the linear model settings, it concurs with AIC for the logistic models where less information is provided from the binary responses.

In the first sampling plan, the inclusion probabilities are related to Y only through a single covariate in the model (*i.e.*, X_6 or X_7). The sample correlation structure between the response and covariates is largely maintained from the finite population. Consequently, no substantial difference is observed between the weighted and unweighted selection procedures from Table 5.1.

The insights of using sampling weights in variable selection are tentatively revealed from the second sampling plan, where the sample correlation structure is systemically distorted. Clearly, the spurious correlation between covariates in the sampled units deteriorates the efficiency of selection methods. This is reflected from the depressed PSRs and the inflated FDRs from the unweighted procedures. Incorporating sampling weights in the selecting process is helpful to correct the biased result. In particular, noticeable improvements have been observed for the BIC-based selection. In the most impressive case (*i.e.*, Model 3 of Table 5.2), the pseudo-likelihood-based BIC substantially improves the classic BIC by increasing the PSR from 0.65 up to 0.89, while reduces the corresponding FDR from 0.62 down to 0.50. Our observation echoes the rationale of weighting as the removal of bias due to the informative sampling (Section 6.3, Fuller 2009).

Table 5.1
Selection for the design not generating strong spurious correlations (1st plan). Results are summarized in terms of positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR) and averaged model size (AMS); Prediction assessments for Models 1-2 are based on the testing residual sum of squares (RSS), while for Models 3-4 they are based on positive/negative prediction rate (PPR, NPR) with a benchmark 0.5.

Weights	Criterion	PSR	FDR	CSR	AMS	Prediction
Model 1						
Ignored	GCV	0.96	0.19	0.28	4.9	1.04
	AIC	0.99	0.48	0.05	8.7	1.08
	BIC	0.96	0.19	0.28	4.9	1.04
Included	GCV	0.95	0.24	0.19	5.2	1.05
	AIC	0.99	0.61	0.01	11.4	1.11
	BIC	0.95	0.24	0.20	5.3	1.05
Model 2						
Ignored	GCV	0.72	0.19	0.02	5.5	1.07
	AIC	0.89	0.44	0.01	10.3	1.09
	BIC	0.73	0.19	0.03	5.6	1.07
Included	GCV	0.74	0.24	0.02	6.1	1.08
	AIC	0.89	0.54	0.01	12.5	1.12
	BIC	0.74	0.24	0.03	6.1	1.08
Model 3						
Ignored	GCV	0.99	0.59	0.00	7.8	(0.71, 0.45)
	AIC	0.99	0.62	0.00	8.4	(0.69, 0.49)
	BIC	0.96	0.43	0.00	5.1	(0.72, 0.44)
Included	GCV	0.99	0.67	0.00	9.9	(0.71, 0.47)
	AIC	0.99	0.70	0.00	10.7	(0.68, 0.48)
	BIC	0.94	0.45	0.00	5.3	(0.71, 0.45)
Model 4						
Ignored	GCV	0.97	0.44	0.01	9.4	(0.66, 0.55)
	AIC	0.98	0.47	0.01	9.8	(0.65, 0.56)
	BIC	0.87	0.26	0.07	6.0	(0.69, 0.53)
Included	GCV	0.98	0.54	0.01	11.4	(0.66, 0.54)
	AIC	0.98	0.56	0.00	11.9	(0.66, 0.55)
	BIC	0.86	0.30	0.05	6.2	(0.68, 0.53)

Table 5.2

Selection for the design generating strong spurious correlations (2nd plan). Results are summarized in terms of positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR) and averaged model size (AMS); Prediction assessments for Models 1-2 are based on the testing residual sum of squares (RSS), while for Models 3-4 they are based on positive/negative prediction rate (PPR, NPR) with a benchmark 0.5.

Weights	Criterion	PSR	FDR	CSR	AMS	Prediction
			Model 1			
Ignored	GCV	0.83	0.23	0.17	4.6	1.09
	AIC	0.97	0.49	0.04	8.6	1.10
	BIC	0.83	0.23	0.17	4.6	1.09
Included	GCV	0.95	0.31	0.13	5.9	1.07
	AIC	0.99	0.65	0.00	12.5	1.12
	BIC	0.95	0.30	0.14	5.9	1.07
			Model 2			
Ignored	GCV	0.62	0.22	0.02	5.0	1.13
	AIC	0.88	0.45	0.01	10.3	1.14
	BIC	0.62	0.22	0.02	5.1	1.12
Included	GCV	0.72	0.28	0.01	6.5	1.10
	AIC	0.89	0.59	0.00	13.7	1.12
	BIC	0.72	0.27	0.01	6.5	1.10
			Model 3			
Ignored	GCV	0.87	0.62	0.00	7.3	(0.66, 0.44)
	AIC	0.88	0.63	0.00	7.6	(0.65, 0.45)
	BIC	0.65	0.62	0.00	4.5	(0.68, 0.42)
Included	GCV	0.97	0.74	0.00	11.9	(0.70, 0.46)
	AIC	0.97	0.75	0.00	12.4	(0.68, 0.46)
	BIC	0.89	0.50	0.00	5.6	(0.70, 0.44)
			Model 4			
Ignored	GCV	0.94	0.48	0.00	9.5	(0.62, 0.51)
	AIC	0.95	0.50	0.00	10.0	(0.62, 0.52)
	BIC	0.72	0.41	0.00	6.1	(0.64, 0.49)
Included	GCV	0.93	0.61	0.00	12.5	(0.64, 0.53)
	AIC	0.94	0.62	0.00	12.9	(0.64, 0.53)
	BIC	0.82	0.34	0.01	6.4	(0.67, 0.54)

5.3 Scenario 2: Model mis-specification

A well-known rationale for using sampling weights is that it provides protection against model mis-specification (Pfeffermann and Holmes 1985; Kott 1991): the inferences based on weighted estimates may remain valid for the surveyed population, even when the model fails. To gain further insights of weighting in variable selection, we further compare the proposed BIC with the classic unweighted methods in the simulation where the presumed model is misspecified from the model that generates the data. In such situations, a postulated “true” model does not exist, and the goal of variable selection is to find an optimal model that well describes the finite population. We still make use of the stratified pseudo-finite population in Section 5.2, but generate the response variable Y according to the strata. Specifically, the values of Y for units in strata (Male, 55+) and (Female, 55+) were generated by

$$Y = 0.6X_6 + 0.4X_{18} + 0.4X_{20} + 0.6X_{38} + \varepsilon,$$

while the values Y for units in the strata (Male, 55-) and (Female, 55-) were generated by

$$Y = 0.6X_6 + 0.4X_{18} + 0.4X_{20} + \varepsilon$$

with $\varepsilon \sim N(0, 1)$ denoting a random error. In other words, we assume that variable X_{38} is influential only for people aged 55 and older, but not for people younger than 55. In addition, we further violate the presumed Model 1 by excluding X_6 from the set of candidate covariates, which mimics the situation where one important design feature is omitted in the modeling. A stratified SRSWOR of size 500 or 1,000 is drawn using the first sampling plan in Section 5.2. The weighted and unweighted procedures are then tested for the variable selection based on the sampled units.

We summarize the simulation results in Table 5.3 by estimating the selection rates of X_{18} , X_{20} , and X_{38} based on 1,000 replications. Similar to the previous simulations, the averaged model size (AMS) and the testing RSS of selected models (*i.e.*, the averaged RSS based on testing data of size 200) are also included in the summary. From Table 5.3, we see that when the model assumption is violated, the pseudo-likelihood-based BIC still achieves relatively high prediction accuracy by suggesting relevant variables with high probability. In contrast, ignoring the survey weights leads to nearly 9% relative loss on the testing RSS because of the exclusion of X_{38} . Apparently, increasing the sample size helps to improve the goodness of fit for the misspecified models, yet the improvement is at a cost by including more variables.

Table 5.3
Selection frequency of influential variables in model mis-specified case; The averaged model size (AMS) and the testing residual sum of squares (RSS) are also reported.

Weights	Criterion	X_{18}	X_{20}	X_{38}	AMS	Testing RSS
$n = 500$						
Ignored	GCV	0.78	0.95	0.56	5.9	1.93
	AIC	0.95	0.99	0.73	12.5	1.95
	BIC	0.83	0.97	0.60	6.6	1.93
Included	GCV	0.73	0.92	0.84	6.3	1.77
	AIC	0.91	0.99	0.85	12.5	1.79
	BIC	0.78	0.94	0.83	6.9	1.77
$n = 1,000$						
Ignored	GCV	0.96	1.00	0.79	7.6	1.87
	AIC	0.99	1.00	0.87	13.1	1.88
	BIC	0.97	1.00	0.80	7.9	1.87
Included	GCV	0.93	1.00	0.94	7.6	1.71
	AIC	0.98	1.00	0.96	13.0	1.72
	BIC	0.94	1.00	0.94	7.7	1.71

5.4 Analysis of SLCDC data

To illustrate the application of proposed BIC, we use it to identify health behaviors that affect the control of blood pressure using SLCDC 2009. The response variable is BMHX_02 from the working data obtained from SLCDC, which has 2 levels indicating whether or not the blood pressure of the respondent is under control, based on the latest measurement by a health professional. We treat the remaining 39 variables in the working data as candidate covariates, and our goal is to identify the influential covariates that are associated with blood-pressure control. We build a logistic regression of BMHX_02 on the candidate covariates and use PPL-BIC with SCAD penalty to select the influential ones (weights are scaled by $k = 10^{-3}$). As a preliminary step, each covariate is standardized such that the corresponding first and second weighted sample moments are zero and unity respectively. For comparison, the AIC and GCV are also used in the analysis.

In Figure 5.1, we plot the scores of criterion with respect to the degree of model sparsity. We see that the BIC selects a model with 11 covariates, while the GCV and AIC pick the same model with 24 covariates. When survey weights are ignored in the selection procedure, models with 7 or 21 covariates are suggested based on the standard BIC or GCV/AIC. The distinction between the weighted and unweighted selection results reflects the potential distortion in the correlation structure of the sampled units. Such a distinction may also be explained by model mis-specification for part of the SLCDC population (Lohr and Liu 1994). Given the potential bias for unweighted methods, the weighted selection results are more plausible in the analysis.

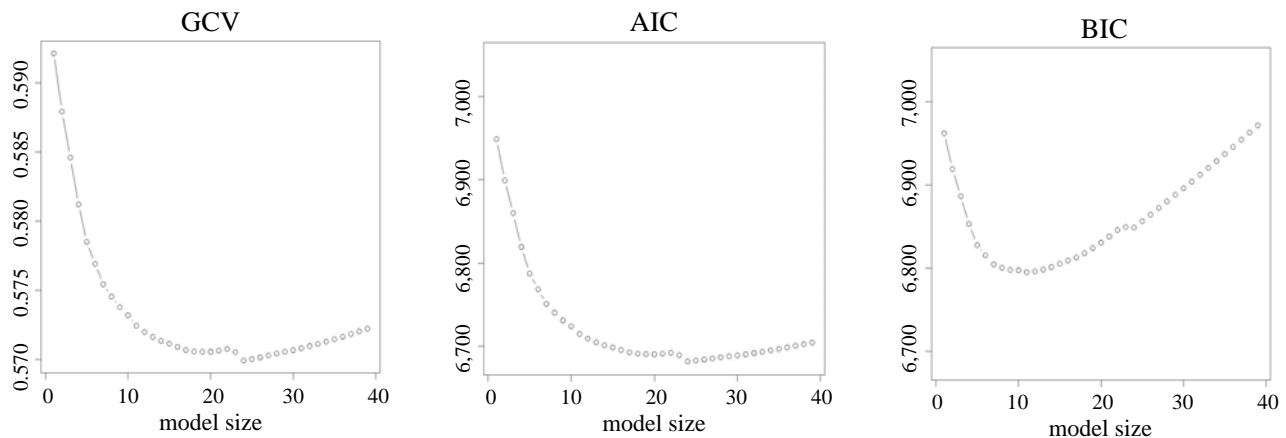


Figure 5.1 Selection criteria values based on candidate models

We further assess the selected models in terms of predictive accuracy as follows. First, we draw 500 independent sets of 5,868 bootstrap samples (with replacement) from the working data of SLCDC. For the t^{th} bootstrap sample d_t , $t = 1, \dots, 500$, the survey weight w_i for the i^{th} unit is adjusted by $\tilde{w}_{ti} = v_{ti} w_i$ with v_{ti} denoting the number of times that the i^{th} unit is selected in d_t . We then fit the selected models to each bootstrap sample (with weights accounted accordingly), and evaluate their weighted positive and negative prediction rates (WPPR, WNPR) by

$$\text{WPPR} = \frac{\sum_{i \notin d_t} w_i I(\hat{y}_i = 1, y_i = 1)}{\sum_{i \notin d_t} w_i I(y_i = 1)}, \quad \text{WNPR} = \frac{\sum_{i \notin d_t} w_i I(\hat{y}_i = 0, y_i = 0)}{\sum_{i \notin d_t} w_i I(y_i = 0)},$$

where y_i and \hat{y}_i denote the i^{th} response in BMHX_02 and its predicted value. We summarize the averaged WPPR and WNPR based on 500 bootstrap samples in Table 5.4 according to three different benchmark values (*i.e.*, 0.25, 0.35, 0.45).

From Table 5.4, we observe that the models selected from unweighted analysis have lower WPPR in general, which provides additional support for using survey weights in the selection procedure. Compared with GCV/AIC, the BIC selects the model with a slightly conservative WPPR but a higher WNPR. Nevertheless, the difference is not significant. Noticeably, the size of BIC-selected model is much less

than the GCV/AIC selected one, which provides an easier interpretation between the response BMHX_02 and the covariates.

Table 5.4
Prediction accuracy for selected models: (WPPR, WNPR) based on different benchmarks.

Weights	Criteria	≥ 0.25	≥ 0.35	≥ 0.45
Ignored	AIC/GCV	(0.646, 0.525)	(0.460, 0.688)	(0.299, 0.811)
	BIC	(0.649, 0.513)	(0.445, 0.705)	(0.265, 0.818)
Included	AIC/GCV	(0.645, 0.523)	(0.488, 0.682)	(0.338, 0.790)
	BIC	(0.654, 0.532)	(0.485, 0.706)	(0.322, 0.830)

To assess the stability of selection, we repeat the weighted selection procedure based on the 500 bootstrap samples. In Table 5.5, we list the bootstrap selection rate for the seven most significant covariates according to their MLE in the original SLCDC working data. The corresponding coefficient estimates and standard errors are also included based on the bootstrap samples. From Table 5.5, we find that only four significant covariates (*i.e.*, DHHX_AGE, GENXDHMH, INHX_06, HWTDBMI) are consistently selected by BIC, while the GCV/AIC tends to pick more unreliable ones in the model. The BIC-based selection result suggests that the control of blood pressure is strongly associated with age, body weights, mental health and the medication information. Our observation echoes many hypertension studies in the literature (see, *e.g.*, Gelber, Gaziano, Manson, Buring and Sesso 2007; Yan, Liu, Matthews, Daviglus, Ferguson and Kiefe 2003).

Table 5.5
Bootstrap selection results for significant variables: (Estimated coefficient, Standard error, Selection rate).

Variable	GCV	AIC	BIC
GEO_ON	(0.14, 0.09, 0.86)	(0.16, 0.09, 0.92)	(0.09, 0.09, 0.58)
DHHX_AGE	(-0.29, 0.09, 1.0)	(-0.32, 0.09, 1.0)	(-0.27, 0.08, 1.0)
GENXDHMH	(-0.15, 0.05, 0.99)	(-0.15, 0.05, 0.99)	(-0.14, 0.06, 0.92)
SMHXDSLTL	(0.11, 0.07, 0.76)	(0.12, 0.07, 0.84)	(0.08, 0.09, 0.47)
MOHXDBPM	(-0.08, 0.07, 0.67)	(-0.09, 0.06, 0.81)	(-0.05, 0.07, 0.35)
INHX_06	(0.18, 0.06, 0.97)	(0.18, 0.06, 0.99)	(0.18, 0.07, 0.91)
HWTDBMI	(0.14, 0.06, 0.95)	(0.14, 0.06, 0.97)	(0.13, 0.06, 0.91)
Ave. Model Size	23.1	27.8	10.3

6 Concluding remarks

In this paper, we have addressed the variable selection problem in the analysis of complex surveys. When units are selected through disproportionate sampling, the data correlation structure reflected in the sample can be distorted. Incorporating sampling weights in the selection process is protective against the biased selection results. In this spirit, we derived a survey-weighted BIC criterion based on the pseudo-likelihood and further proposed an efficient procedure (PPL) for its implementation. With some regularity

conditions, we showed that our criterion consistently identifies the influential variables under a joint randomization framework. The decent performances of proposed method was confirmed by numerical studies.

Acknowledgements

The authors are grateful to the associate editor and the two anonymous referees for their insightful comments and valuable suggestions. The authors are indebted to Professor J.N.K. Rao of Carleton University for his constructive comments to an earlier manuscript. This work was supported by Statistics Canada and MITACS.

Appendix

Table A.1

Variables for analysis of SLCDC data with non-response adjustments: A: allocate to other categories; D: delete from the data; M: impute by mean values; NA: no adjustment applied.

	Variable	Description	Levels	Missing	Adjust
1	BMHX_02	Blood pressure control status	2	1.6%	D
2	GEO_QB	Provinces grouped by region - QC	2	--	NA
3	GEO_ON	Provinces grouped by region - ON	2	--	NA
4	GEO_BC	Provinces grouped by region - BC	2	--	NA
5	GEO_PR	Provinces grouped by region - PR	2	--	NA
6	DHHX_AGE	Age	Cont.	--	NA
7	DHHX_SEX	Sex	2	--	NA
8	GENXDHMH	Perceived mental health	2	0.2%	A
9	CNHX_05	High blood pressure - age when diagnosed	Cont.	2.7%	D
10	MEHX_02	No. of medications taken	Cont.	0.3%	M
11	MEHX_03	No. of times per day medications taken	Cont.	0.1%	M
12	MEHXGMED	No. of medications for high blood pressure	Cont.	2.0%	M
13	MEHX_06	No. of times per day bp medication taken	Cont.	1.0%	M
14	MEHXDMCO	Medication compliance - overall	2	0.2%	A
15	HUHXDHP	Consulted family doctor about hbp	2	0.1%	A
16	SMHX_11A	Smoked at any time since being diagnosed	2	0.1%	A
17	SMHX_13A	Drank alcohol since being diagnosed	2	0.2%	A
18	SMHXDSLTL	Daily salt intake	2	0.2%	A
19	SMHXDFDC	Dietary foods	2	0.1%	A
20	SMHXDPAC	Exercise/physical activity	2	0.1%	A
21	SMHXDBW	Body weight control	2	0.2%	A
22	MOHXDBPM	Self-monitoring of blood pressure	2	0.3%	A
23	MOHX_02	Correct use of bp measurement device	2	0.5%	A
24	INHX_01A	Info from family doctor	2	2.4%	A
25	INHX_01F	Info from family member/friend	2	2.4%	A
26	INHX_02A	Info from book, pamphlet, brochure	2	1.5%	A
27	INHX_02C	Info from package insert with medication	2	1.5%	A
28	INHX_02G	Info from media	2	1.5%	A
29	INHX_02H	Info from internet	2	1.5%	A
30	INHX_04	Info received - emotional impact of hbp	2	0.8%	A
31	INHX_06	Info received - correct use of medication	2	0.6%	A
32	INHX_07	Info received - additional information	2	0.9%	A
33	CPGFGAM	Gambling activity	2	0.5%	A
34	DHHDECF	Household type	2	0.2%	A
35	EDUDH04	Highest level of education in household	2	3.4%	A
36	FVCGTOT	Daily consumption - fruits and vegetables	2	5.2%	A
37	GEODUR2	Urban and rural areas	2	--	NA
38	HWTDBMI	Body mass index (BMI) self-report	Cont.	2.1%	M
39	INCDRPR	Household income - provincial level	10	9.6%	A
40	SACDTOT	Total number hours - sedentary activities	Cont.	1.5%	M

Table A.2

Influential and design variables in simulation settings: * - influential variable to the response; • - design variable affecting sampling probabilities in the 1st plan; ◊ - design variable affecting sampling probabilities in the 2nd plan.

	Variable	Model 1	Model 2	Model 3	Model 4
6	DHHX_AGE	* • ◊	* • ◊	• ◊	• ◊
7	DHHX_SEX	• ◊	• ◊	* • ◊	* • ◊
8	GENXDHMH			* ◊	* ◊
10	MEHX_02	*	*		
18	SMHXDSLTL	* ◊	* ◊		
22	MOHXDBPM	* ◊	* ◊		
26	INHX_02A			* ◊	* ◊
28	INHX_02G				*
30	INHX_04		*		
34	DHHDECF		*		
36	FVCGTOT				*

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (Eds., B.N. Petrox and F. Caski), 267-281.
- Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Binder, D., and Roberts, G. (2003). *Analysis of Survey Data*, Chapter: Design-based and model-based methods for estimating model parameters. Wiley Series in Survey Methodology, Chichester.
- Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 377-403.
- Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.
- Frank, I.E., and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.
- Fuller, W.A. (2009). *Sampling Statistics*. Wiley, Hoboken.
- Gelber, R.P., Gaziano, J.M., Manson, J.E., Buring, J.E. and Sesso, H.D. (2007). A prospective study of body mass index and the risk of developing hypertension in men. *American Journal of Hypertension*, 20, 370-377.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54, 127-138.

- Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, 175-188.
- Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.
- Kott, P.S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, 45, 107-112.
- Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21, 1225-1248.
- Lohr, S.L., and Liu, J. (1994). A comparison of weighted and unweighted analyses in the NCVS. *Journal of Quantitative Criminology*, 10, 343-360.
- Mallows, C.L. (1973). Some Comments on C_p . *Technometrics*, 15, 661-675.
- Molina, E.A., and Skinner, C.J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis*, 13, 395-405.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- Pfeffermann, D., and Holmes, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A*, 148, 268-278.
- Rahiala, M., and Teräsvirta, T. (1993). Business survey data in forecasting the output of Swedish and Finnish metal and engineering industries: A Kalman filter approach. *Journal of Forecasting*, 12, 255-271.
- Royall, M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- She, Y. (2011). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, in press.
- Skinner, C. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics*, manuscript.
- Statistics Canada (2009). Survey on living with chronic diseases in Canada 2009: User guide. Supplementary documentation.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 111-147.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

- Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.
- Wolfson, W.G. (2004). Analysis of labour force survey data for the information technology occupations 2000-2003. *Report for the Software Human Resource Council*, WGW Services Ltd., Ottawa, Ontario.
- Xie, B., Pan, W. and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64, 921-930.
- Xu, C., and Chen, J. (2012). Technical supplement to “Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data”. Available from the first author.
- Yan, L.L., Liu, K., Matthews, K.A., Daviglius, M., Ferguson, T.F. and Kiefe, C.I. (2003). Psychosocial factors and risk of hypertension: The coronary artery risk development in young adults (CARDIA) study. *The Journal of the American Medical Association*, 290, 2138-2148.