Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

SEMIPARAMETRIC REGRESSION TECHNIQUE FOR COMPLEX SURVEY DATA

Zilin Wang and David Bellhouse[1]

ABSTRACT

Due to the usual complexity of the sampling designs, survey data are neither independent nor identically distributed. Hence, estimation methods developed for independent and identically distributed data are not suitable for models based on complex survey data. The aim of this paper is to discuss such estimation methods with emphasis on semiparametric regression models for complex survey data, in which explanatory variables are composed of nonparametric and parametric parts. Estimation methods for this model combine the nonparametric local polynomial regression estimation and the classic least squares estimation in complex surveys. Moments and asymptotic properties of the estimators are discussed. Methodology and theoretical results will be illustrated using the 1990 Ontario Health Survey.

KEYWORDS: Binning; Ontario Health Survey; Regression; Sampling; Shift Function; Smoothing.

1. INTRODUCTION

Given $\mathbf{y}$ is the response variable and $\mathbf{X}$ is a matrix of corresponding the explanatory variables, a conventional regression model is of the form:

$$E(\mathbf{y}|\mathbf{X}) = G(\mathbf{X}),$$

where G(.) is called a regression function. Based on the assumption of the regression function, regression modelling can be divided into two major streams: parametric and nonparametric. A parametric regression is parameterized by an unknown *p*-dimensional parameter vector, $\boldsymbol{\beta}$ and $G(\mathbf{X})$ is usually denoted as $G(\mathbf{X}, \boldsymbol{\beta})$. We estimate $\boldsymbol{\beta}$ with the assumption on the functional form of G(. , .). If the assumption for the form of G(. , .) is correct, the performance of the parametric regression model is very useful, however, once misspecification occurs, misleading results can be obtained. In a nonparametric regression model, we relax the assumption on the form of G(.) and use the local information to obtain the point estimates of the function G(.). A nonparametric regression model can be estimated by a smoother. Even though nonparametric regression techniques have demonstrated their usefulness, when we conduct flexible regression modeling we pay a price for relaxing the assumption of a specific functional form in nonparametric regression analysis. In particular, beyond the difficulty of choosing the right window size (neighbourhood), a more serious problem that relates to all smoothing methods for a multiple regression model is the "curse of dimensionality", which happens when neighborhoods with a fixed number of points become less local as the number of dimensions increases. The "curse of dimensionality" makes the rate of convergence of an estimator so slow that the performance of nonparametric estimation for multiple regressions is not promising. One result of the "curse of dimensionality" is the infeasibility of including discrete explanatory variables in the nonparametric regression analysis.

To take advantage of the strength of parametric estimation and to minimize the occurrence of "curse of dimensionality", a so-called partial linear semi-parametric regression model is devised of the form,

$$E(\mathbf{y}|\mathbf{X},\mathbf{z}) = \mathbf{X}\boldsymbol{\beta} + G(\mathbf{z}),$$

where the explanatory variables are represented separately in two parts: the nonparametric part ($G(\mathbf{z})$) and the parametric linear part ($\mathbf{X}\boldsymbol{\beta}$). In this semi-parametric regression model, both the functional form of the

[1] Zilin Wang and David Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada, N6A 5B7.

nonparametric part of the model and the parameters will be estimated. This partial linear semiparametric model has a priori motivation as a data analytic tool and retains an important interpretive feature. We can put those variables with more known information on the functional form in the parametric part of the model and the variable with little information on the functional form in the nonparametric part of the model. In addition, discrete explanatory variables have always created problems in nonparametric regression estimation because of low effective sample sizes. It is very natural to include the discrete explanatory variables in the linear part of the model.

The objective of this paper is to apply this partial linear semi-parametric regression model to complex surveys. We are interested in the estimating procedures developed independently by Robinson (1988) and Speckman (1988) for independent and identically distributed data. Due to the sampling design, data from a complex survey are neither independent nor identically distributed. Hence, we cannot directly apply the estimation method of Robinson (1988) and Speckman (1988) to complex survey data. To solve the technical difficulty of complex data, we establish two superpopulations such that we can adapt the estimation method for independent and identically distributed data and make inferences for the survey sample estimators.

Due to Robinson (1988) and Speckman (1988), the estimation procedure for partial linear regression models consists of a nonparametric estimation method and a least squares estimation method. As a result, we need a smoother to accomplish the estimating procedure in the sampling context. The smoother we will use is developed by Bellhouse and Stafford (2001) for complex surveys. One of the characteristics of complex survey data is that the size of the data set can be very large. Usually, there are multiple observations at distinct values in a large survey data set. Large-scale data sets not only can result in non-informative trends between the response variable and the covariates when plotting the data, they also make the estimation process very computationally cumbersome. Hence, it is very natural in the complex survey data analysis to bin the data into domains according to the distinct values of the characteristic variables. In Bellhouse and Stafford (2001), local polynomial regression methods are put forward for large-scale surveys and rely on binning the data on the explanatory variable.

Combining the well-established least squares estimation technique and the local polynomial regression techniques developed by Bellhouse and Stafford (2001) for complex surveys, we develop the survey sample estimators for the partial linear regression model and establish their asymptotic properties. The paper is organized as follows. In section 2, we introduce the partial linear regression model in the sampling context. In section 3, asymptotic properties of the survey estimators are discussed. An empirical illustration of the estimation method using the 1990 Ontario Health Survey in are carried out in Section 4. The paper is concluded in section 5.

2. A PARTIAL LINEAR MODEL IN THE SAMPLING CONTEXT

2.1 Preliminary

A semiparametric model is defined as:

$$\mathbf{y} = \mathrm{G}(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is the vector of response variable and $\boldsymbol{\varepsilon}$ is independent and identically distributed with mean zero and constant variance. G(.) is an arbitrary function of $\mathbf{z}$. Based on the model information on the independent variables, the independent variables are separated into two types. Independent variables included in the $n \times p$ matrix $\mathbf{X}$ correspond to the parametric or linear part of the model and independent variable, $\mathbf{z}$, is the nonparametric part of the model. Each parametric independent variable, $\mathbf{x}_j$, is a vector of random variables with distribution F_j. $\mathbf{z}$ is measured on a continuous scale and $\mathbf{X}$ contains either continuous or discrete explanatory variables. Both the functional form of G(.) and parameters $\beta_1, \ldots, \beta_p$ are unknown. Additionally, it is assumed that $\mathrm{E}(\boldsymbol{\varepsilon} | \mathbf{z}, \mathbf{X}) = \mathbf{0}$ and that there are no interactions between $\mathbf{X}$ and $\mathbf{z}$.

The problem in estimating $\boldsymbol{\beta}$ in the partial linear model as stated in (1) is that there is a function of unknown form, $\mathrm{G}(\mathbf{z})$. If it were possible to find a way to remove this function, the least squares procedure can be used to estimate the resulting linear regression model.

Taking the expectation of both sides of equation (1) conditional on $\mathbf{z}$ yields,

$$\mathrm{E}(\mathbf{y} \mid \mathbf{z}) = \mathrm{E}(\mathbf{X} \mid \mathbf{z})\boldsymbol{\beta} + \mathrm{G}(\mathbf{z}) \tag{2}$$

given that $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{z}) = \mathbf{0}$. Now, subtract (2) from (1) to obtain

$$\mathbf{y} - \mathrm{E}(\mathbf{y} \mid \mathbf{z}) = (\mathbf{X} - \mathrm{E}(\mathbf{X} \mid \mathbf{z}))\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

By defining $\boldsymbol{Y} \equiv \mathbf{y} - \mathrm{E}(\mathbf{y}|\mathbf{z})$ and $\boldsymbol{X} \equiv \mathbf{X} - \mathrm{E}(\mathbf{X}|\mathbf{z})$, we get the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

Now one obvious approach is to estimate $\boldsymbol{\beta}$ by the method of least squares. Unfortunately, since $\mathrm{E}(\mathbf{y}|\mathbf{z})$ and $\mathrm{E}(\mathbf{X}|\mathbf{z})$ are unknown, least squares estimation of $\boldsymbol{\beta}$ is not feasible. Consequently, we carry out the estimation of $\boldsymbol{\beta}$ in two steps. In the first step, the conditional expectations appearing in equation (3) is estimated with Nadaya-Watson kernel smoothing technique. In the second step, $\mathrm{E}(\mathbf{y}|\mathbf{z})$ and $\mathrm{E}(\mathbf{X}|\mathbf{z})$ in (3) are replaced with their estimates obtained in the first step and estimate $\boldsymbol{\beta}$ with the method of least squares.

Once the estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is obtained, the difference between the response variable $\mathbf{y}$ and the $\mathbf{X}\hat{\boldsymbol{\beta}}$ is treated as the dependent random variable and function $\mathrm{G}(\cdot)$ is estimated in accordance with the following model,

$$\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \quad \mathrm{G}(\mathbf{z}) + \boldsymbol{\mu} \tag{5}$$

The advantage of this semiparametric method for independent and identically distributed data is that iteration is not required and root *n* consistencies of the estimators of the linear coefficients can be achieved.

2.2 Sampling Design and Superpopulations

Suppose that we have a population $\boldsymbol{U}$ consisting of N distinct units. The characteristic of interest is a vector valued unit $(y_k, \boldsymbol{x}_k, z_k)$ for all the $k=1,\ldots,N$. y_k represents the k^{th} population value of the response variable and $(\boldsymbol{x}_k, z_k)$ represents the k^{th} observation of the explanatory variables and is a vector with length $p+1$. Let $\mathbf{s}$ be a set of units in the sample with $(y_k, \boldsymbol{x}_k, z_k, w_k)$ for $k \in \boldsymbol{s}$ obtained according to the sampling design with sample size n. The survey weight w_k is attached to the k^{th} sampling unit. Additionally, we assume that there is zero nonresponse to assure that the inclusion probabilities are equal to the reciprocal of the sampling weights.

Note that there are several estimation procedures needed to accomplish estimation for the partial linear model. In order to obtain estimates and make inferences with them, we need to assume a superpopulation framework. What is typically used in survey data analysis is to assume a working model or a superpopulation model on the finite population. The parameter estimates of this model yield finite population parameters or census estimates based on the model. The survey sample is used to obtain estimates of these census "estimates." Asymptotic derivations to justify inferences from the sample to the population are normally obtained through a second superpopulation model.

Superpopulation 1

The N finite population units are a sample of independent and identically distributed units from the infinite superpopulation. The units of a finite population are realization of the model defined in equation (1). We denote $\mathbf{B} = (\mathrm{B}_1, \ldots, \mathrm{B}_p)$ and $\mathrm{g}(z)$ as the finite population parameters of the linear coefficients and the regression function at fixed point z in the working model, respectively. Based on this superpopulation, we can directly adopt the methodology for the independent and identical distributed data to obtain finite population parameters of the interest that we can estimate. Our concern is only that the finite population parameters are consistent estimators of the superpopulation's parameters when the assumption of independence is withdrawn. Superpopulation 1 does not only allow us to derive valid asymptotic results in the view of independence, it may also create specification problems if the finite population units do not agree with the superpopulation model. Hence, once sample estimators are obtained, another

superpopulation, which is composed of a nested sequence of finite populations, is used to establish the asymptotic distribution and inferences from the estimators.

Superpopulation 2

Superpopulation 2 consists of a nested sequence of finite populations indexed by υ such that all the finite population quantities and the sample quantities depend on the index υ and all the asymptotic frameworks are established as $\upsilon \to \infty$.

2.3. Estimation

Using the framework of Superpopulation 1, we extend the estimation procedure in Section 2.1 to the complex survey data with some modifications. Specifically, instead of the Nadaya-Watson kernel smoothing technique used in Robinson (1988), we will use the local polynomial regression technique to estimate the conditional expectations. As mentioned in Wand and Jones (1995), the Nadaya-Watson kernel smoothing technique can be considered as local constant fitting and it has been shown to have higher boundary bias than some other degrees of local polynomial regression fits. It is noted that estimation of the linear coefficients and the nonparametric regression function are accomplished in two steps. In the first step, we estimate the linear coefficients and in the second step we estimate the nonparametric function.

When conducting the first step in the estimation procedure, a smoother is needed to estimate the conditional expectations of the response variable and the parametric explanatory variables on the nonparametric explanatory variable, **z**. We denote $m_{\mathbf{y}}(z)$ and $m_{\mathbf{x}_j}(z)$ the population conditional expectations of $\mathbf{y}$ and $\mathbf{x}_j$ on a fixed point z, respectively. In order to estimate $m_{\mathbf{y}}(z)$ and $m_{\mathbf{x}_j}(z)$, we bin the observed data according to **z**. Suppose that **z** has m distinct values in the finite population. Let z_i denote the i^{th} distinct value or the i^{th} bin and assume that the values of z_i are equally spaced with length $z_i - z_{i-1}$. The finite population proportion of the observations with z_i is denoted by p_i. Let the vector of finite population means for response variable y at distinct values of z be $\bar{\mathbf{y}} = (\bar{y}_1, \cdots, \bar{y}_m)$ and the vector of finite population means for the j^{th} independent variable $\mathbf{x}_j$ for $j = 1,\ldots, p$ be $\bar{\mathbf{x}}_j = (\bar{x}_{1j}, \cdots, \bar{x}_{mj})$. $\hat{\bar{y}}_i$, $\hat{\bar{x}}_{ij}$ and $\hat{\bar{p}}_i$ are the survey estimators of $\bar{y}_i$, $\bar{x}_{ij}$ and $\bar{p}_i$ for all $i = 1, \ldots, m$, respectively.

Based on the binned sample means and sample proportion, we have the survey estimator of $m_{\mathbf{y}}(z)$ and $m_{\mathbf{x}_j}(z)$ at z_i for all $i = 1, \ldots, m$ as,

$$\hat{\mathrm{E}}(\mathbf{x}_j \mid z = z_i) = \hat{\mathrm{m}}_{\mathbf{x}_j}(z_i) = \mathbf{e}^T (\mathbf{Z}_{z_i}^T \hat{\mathbf{K}}\mathbf{W}_{z_i} \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^T \hat{\mathbf{K}}\mathbf{W}_{z_i} \hat{\bar{\mathbf{x}}}_j, \tag{6}$$

and

$$\hat{\mathrm{E}}(\mathbf{y} \mid z = z_i) = \hat{\mathrm{m}}_{\mathbf{y}}(z_i) = \mathbf{e}^{\mathrm{T}} (\mathbf{Z}_{z_i}^{\mathrm{T}} \hat{\mathbf{K}}\mathbf{W}_{z_i} \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^{\mathrm{T}} \hat{\mathbf{K}}\mathbf{W}_{z_i} \hat{\bar{\mathbf{y}}}, \tag{7}$$

where

$$\mathbf{Z}_{z_i} = \begin{pmatrix} 1 & z_1 - z_i & \cdots & (z_1 - z_i)^q \\ 1 & z_2 - z_i & \ldots & (z_2 - z_i)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_m - z_i & \cdots & (z_m - z_i)^q \end{pmatrix}$$

and

$$\hat{\mathbf{KW}}_{z_i} = \frac{1}{h}\boldsymbol{diag}(\hat{p}_1 K(\frac{z_1 - z_i}{h}),\ldots,\hat{p}_m K(\frac{z_m - z_i}{h}))$$

$K(\cdot)$ is a kernel function satisfying $\int K(t)\,dt = 1$ and $\int K(t)^2\,dt < \infty$ and q is the degree of polynomial the model fits. h is the bandwidth that controls the size of neighbourhood. The vector $\mathbf{e}$ is the $m \times 1$ vector of the form $(1,0, \ldots, 0)^{\mathrm{T}}$. The vector $\hat{\bar{\mathbf{y}}}$ is the $m \times 1$ vector of the form $(\hat{\bar{y}}_1, \ldots, \hat{\bar{y}}_m)^{\mathrm{T}}$, and $\hat{\bar{\mathbf{x}}}_j$ is the $m \times 1$ matrix of the form $(\hat{\bar{x}}_{1j},\cdots,\hat{\bar{x}}_{mj})^{\mathrm{T}}$.

In order to estimate the finite population $\mathbf{B}$, we need to reconstruct the data in such a way that the working model shown in equation (3) can be used. Let N_i be the number of observations that fall in the i^{th} bin and $\sum_{i=1}^{m} N_i = N$. $\mathbf{M}_{\mathbf{X}}$ is a $N \times p$ matrix consisting of all the population conditional expectations of $\mathbf{X}$ and of the form,

$$\mathbf{M}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} \mathrm{m}_{\mathbf{x}_1}(z_1) & \mathrm{m}_{\mathbf{x}_2}(z_1) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{m}_{\mathbf{x}_1}(z_1) & \mathrm{m}_{\mathbf{x}_2}(z_1) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_1) \end{pmatrix}_{N_1 \times p} \\ \vdots \quad\quad\quad\quad \vdots \\ \begin{pmatrix} \mathrm{m}_{\mathbf{x}_1}(z_m) & \mathrm{m}_{\mathbf{x}_2}(z_m) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_m) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{m}_{\mathbf{x}_1}(z_m) & \mathrm{m}_{\mathbf{x}_2}(z_m) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_m) \end{pmatrix}_{N_m \times p} \end{pmatrix} \tag{8}$$

Similarly, we set up a $N \times 1$ vector, $\mathbf{M}_{\mathbf{y}}$, such that $\mathrm{m}_{\mathbf{y}}(z_i)$ is repeated for N_i times in the i^{th} bin,

$$\mathbf{M}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \mathrm{m}_{\mathbf{y}}(z_1) \\ \vdots \\ \mathrm{m}_{\mathbf{y}}(z_1) \end{pmatrix}_{N_1 \times 1} \\ \vdots \\ \begin{pmatrix} \mathrm{m}_{\mathbf{y}}(z_m) \\ \vdots \\ \mathrm{m}_{\mathbf{y}}(z_m) \end{pmatrix}_{N_m \times 1} \end{pmatrix} \tag{9}$$

Let $\mathbf{y} = (y_1, \ldots, y_N)$, $\mathbf{x}_j = (x_{j1}, \ldots, x_{jN})$ and $\mathbf{X}=(\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Using $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$, we have $\boldsymbol{Y} \equiv \mathbf{y} - \mathbf{M}_{\mathbf{y}}$ and $\boldsymbol{X} \equiv \mathbf{X} - \mathbf{M}_{\mathbf{X}}$. With these transformed data and the working model in equation (3)**,** we have the multiple regression least squares census estimates without constant term:

$$\mathbf{B} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}.$$

Following the population counterpart, we represent the sampling data as (**X**, **y, z,** ***w***) with sample size n and n_i observations within each bin such that $\sum_{i=1}^{m} n_i = n$. $\mathbf{y}$ is a $n \times 1$ vector and of the form $(y_1, \ldots, y_n)$ and $\mathbf{x}$ is a $n \times p$ matrix and of the form $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. We can construct $\hat{\mathbf{M}}_{\mathbf{X}}$ and $\hat{\mathbf{M}}_{\mathbf{y}}$ with the same way as we construct $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$ in equations (8) and (9). That is, we use sampling estimates $\hat{\mathrm{m}}_{\mathbf{x}_j}(z_i)$ and $\hat{\mathrm{m}}_{\mathbf{y}}(z_i)$ that are shown in (6) and (7) to obtain,

$$\hat{\mathbf{M}}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{x}_1}(z_1) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_1) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\mathrm{m}}_{\mathbf{x}_1}(z_1) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_1) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_1) \end{pmatrix}_{N_1 \times p} \\ \vdots \\ \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{x}_1}(z_m) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_m) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_m) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\mathrm{m}}_{\mathbf{x}_1}(z_m) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_m) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_m) \end{pmatrix}_{N_m \times p} \end{pmatrix} \text{ and } \hat{\mathbf{M}}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{y}}(z_1) \\ \vdots \\ \hat{\mathrm{m}}_{\mathbf{y}}(z_1) \end{pmatrix}_{N_1 \times 1} \\ \vdots \\ \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{y}}(z_m) \\ \vdots \\ \hat{\mathrm{m}}_{\mathbf{y}}(z_m) \end{pmatrix}_{N_m \times 1} \end{pmatrix}$$

Defining $\hat{\boldsymbol{Y}} = \mathbf{y} - \hat{\mathbf{M}}_{\mathbf{y}}$ and $\hat{\boldsymbol{X}} = \mathbf{X} - \hat{\mathbf{M}}_{\mathbf{X}}$, we have the estimator of $\mathbf{B}$ in the context of the complex survey:

$$\hat{\mathbf{B}} = (\hat{\boldsymbol{X}}^{\mathrm{T}}\, \mathbf{W}\, \hat{\boldsymbol{X}})^{-1}\, \hat{\boldsymbol{X}}^{\mathrm{T}}\, \mathbf{W} \hat{\boldsymbol{Y}}.$$

where $\mathbf{W}$ is the $n \times n$ weight matrix with design weight, w_k, on the diagonal entry.

Once we obtain the sampling estimator, $\hat{\mathbf{B}}$, we can estimate the population parameter g(·) by taking equation (5) as the working model. By applying the local polynomial technique again and using the sampling estimates $\hat{\mathbf{B}}$, we have

$$\hat{\mathrm{g}}(z_i) = \mathbf{e}^{\boldsymbol{T}} (\mathbf{Z}_{z_i}^{\boldsymbol{T}}\, \hat{\mathbf{K}}\mathbf{W}_{z_i}\, \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^{\boldsymbol{T}}\, \hat{\mathbf{K}}\mathbf{W}_{z_i}\, \hat{\bar{\mathbf{R}}}, \tag{10}$$

where $\hat{\mathbf{R}} = \mathbf{y} - \mathbf{x}\hat{\mathbf{B}}$ and $\hat{\bar{\mathbf{R}}}$ is the vector of binned means of survey estimate $\hat{\mathbf{R}}$.

3. ASYMPTOTIC PROPERTIES

Let $\boldsymbol{\theta}^T = (\mathbf{B}^T{}_{1\times p}, \mathbf{m}_{\mathbf{x}}(\mathbf{z})^T{}_{1\times pm}, \mathbf{m}_{\mathbf{y}}(\mathbf{z})^T{}_{1\times m})$ be a vector of size 1× (p+(p+1)m) and containing all the finite population parameters. Following Binder (1983) and basing the working model in equation (3), we can express the finite population parameters in a normal equation in the following fashion,

$$\mathbf{u}(\boldsymbol{\theta}) \equiv \sum_{k=1}^{N} (\mathbf{x}_k - \mathbf{M}_{\mathbf{x}k})^{\boldsymbol{T}} (y_k - \mathrm{M}_{\mathbf{y}_k}) - \sum_{i=1}^{N} (\mathbf{x}_k - \mathbf{M}_{\mathbf{x}\boldsymbol{k}})^{\boldsymbol{T}} (\mathbf{x}_k - \mathbf{M}_{\mathbf{x}\boldsymbol{k}})\mathbf{B} = \mathbf{0}_{p\times 1} \tag{11}$$

where $\mathbf{M}_{\mathbf{X}k}$ is the k^{th} row of the $N \times p$ matrix $\mathbf{M}_{\mathbf{X}}$ and $\mathrm{M}_{\mathbf{y}k}$ is the k^{th} elements of the $N \times 1$ vector $\mathbf{M}_{\mathbf{y}}$. Both $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$ are defined in equations (8) and (9), of which matrices consist of all the estimated conditional expectations on $\mathbf{z}$. The objective of setting up equation (11) is to obtain the solution of $\mathbf{B}$, which is the least squares estimator of the superpopulation regression model in equation (2).

Analogous to the population normal equation, the survey estimates of $\mathbf{u}(\boldsymbol{\theta})$ is

$$\hat{\mathbf{u}}(\mathbf{B}, \hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z})) \equiv \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{X}k})^{\boldsymbol{T}} (y_k - \hat{\mathrm{M}}_{\mathbf{y}_k}) w_k - \sum_{\boldsymbol{k}\in\mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{X}\boldsymbol{k}})^{\boldsymbol{T}} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{X}\boldsymbol{k}}) \mathbf{B} w_k,$$

where $\hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z})$ is a vector whose elements are the estimated conditional expectations of $\mathbf{y}$ on all the distinct point of $\mathbf{z}$ and $\hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z})$ is a vector of the form $(\hat{\mathbf{m}}_{\mathbf{x}_1}(\mathbf{z}), \ldots, \hat{\mathbf{m}}_{\mathbf{x}_p}(\mathbf{z}))$, where each $\hat{\mathbf{m}}_{\mathbf{x}_j}(\mathbf{z})$ is composed of estimated conditional expectations of $\mathbf{x}_j$ on all the distinct point of $\mathbf{z}$. Given that $\hat{\mathbf{B}}$ is the survey least squares estimator of $\mathbf{B}$ and $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{B}}, \hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z}))$, we have,

$$\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) \equiv \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k})^T (y_k - \hat{\mathrm{M}}_{y_k}) w_k - \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k})^T (\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k}) \hat{\mathbf{B}} w_k = \mathbf{0}_{p\times 1}$$

Taking a Taylor expansion of $\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}})$ at $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, we obtain,

$$\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) \equiv \mathbf{0}_{p\times 1} \approx \hat{\mathbf{u}}(\boldsymbol{\theta}) + \frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}} (\hat{\mathbf{B}} - \mathbf{B}) + \hat{\mathbf{U}}_{\boldsymbol{\xi}} (\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})) \tag{12}$$

where $\hat{\mathbf{m}}_{\xi}(\mathbf{z})$ and $\mathbf{m}_{\xi}(\mathbf{z})$ are two vectors in the form of $(\mathbf{m}_{\mathbf{y}}(\mathbf{z}), \mathbf{m}_{\mathbf{X}}(\mathbf{z}))$ and $(\hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z}))$ respectively. $\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ is a $p \times (p+1)m$ matrix whose components are the first derivatives of $\hat{\mathbf{u}}(\boldsymbol{\theta})$ with respect to $\mathrm{m}_{\mathrm{y}}(z_i)$ and $\mathrm{m}_{\mathbf{x}_j}(z_i)$ for all j = 1, …, p and i = 1, …, m. Note that $\boldsymbol{\xi}$ in the model of interest represents $\mathbf{y}$ or a covariate $\mathbf{x}_j$. Rearranging equation (12), we have,

$$\hat{\mathbf{u}}(\boldsymbol{\theta}) + \hat{\mathbf{U}}_{\boldsymbol{\xi}} (\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})) \approx -\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}} (\hat{\mathbf{B}} - \mathbf{B}) .$$

Taking variances of both sides, we obtain in the limit,

$$\Omega = \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right) V(\hat{\mathbf{B}}) \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right)^T$$

where

$$\Omega = \mathrm{V}(\hat{\mathrm{u}}(\boldsymbol{\theta})) + \mathbf{U}_{\boldsymbol{\xi}}\, \mathrm{V}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}))\mathbf{U}_{\boldsymbol{\xi}}^T + 2\,\mathrm{COV}(\hat{\mathrm{u}}(\boldsymbol{\theta}), \hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}))\mathbf{U}_{\boldsymbol{\xi}}^T .$$

Providing that the matrix $\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}$ is of full rank, we obtain the variance of the sampling estimator of linear coefficients to be

$$V(\hat{\mathbf{B}}) = \left(\frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right)^{-1} \Omega \left(\frac{\partial \mathbf{u}(\boldsymbol{\theta})^T}{\partial \mathbf{B}}\right)^{-1} . \tag{13}$$

Intuitively, equation (13) suggests that the variability of $\hat{\mathbf{B}}$ is caused by the estimated conditional expectations, the survey total from the estimating equation and the covariance between the survey total and the estimated conditional expectations. Given that

$$\hat{\Omega} = \hat{\mathrm{V}}(\hat{\mathrm{u}}(\boldsymbol{\theta})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}\, \hat{\mathrm{V}}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}))\hat{\mathbf{U}}_{\boldsymbol{\xi}}^T + 2\,\hat{\mathrm{COV}}(\hat{\mathrm{u}}(\boldsymbol{\theta}), \hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}))\hat{\mathbf{U}}_{\boldsymbol{\xi}}^T ,$$

an estimator of $\mathrm{V}(\hat{\mathbf{B}})$ is

$$\hat{\mathrm{V}}(\hat{\mathbf{B}}) = \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right)^{-1} \hat{\Omega} \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})^T}{\partial \mathbf{B}}\right)^{-1} .$$

Note that all the estimates of derivatives in equation (12) are of the form of a survey total and domain totals. In addition, the survey estimates of the conditional expectations are dependent of the domain means and the domain proportions. Consequently, **B** defined in equation (12) is merely a function of survey means and domain means. Hence, under superpopulation 2 framework and some regularity conditions stated in the literature (for example, Shao (1998) and Krewski & Rao (1981)), we state the following asymptotic properties of $\hat{\mathbf{B}}$.

Theorem 1:
Under regularity conditions and a multi-stage sampling design, the asymptotic distribution of $\sqrt{n}(\hat{\mathbf{B}}-\mathbf{B})$ is normal with mean zero and variance $V(\hat{\mathbf{B}})$ and $\hat{V}(\hat{\mathbf{B}})$ converges to $V(\hat{\mathbf{B}})$ in probability.

If we define $\hat{\mathbf{A}}_i = \mathbf{e}^T(\mathbf{Z}_{z_i}^T \hat{\mathbf{K}}\mathbf{W}_{z_i}\mathbf{Z}_{z_i})^{-1}\mathbf{Z}_{z_i}^T \hat{\mathbf{K}}\mathbf{W}_{z_i}$, equation (10) can be rewritten as $\hat{g}(z_i) = \hat{\mathbf{A}}_i\hat{\bar{\mathbf{R}}}$. Naturally, $\hat{\mathbf{A}}_i$ and $\hat{\bar{\mathbf{R}}}$ are the survey estimators of $\mathbf{A_i} \equiv \mathbf{e}^T(\mathbf{Z}_{z_i}^T\mathbf{K}\mathbf{W}_{z_i}\mathbf{Z}_{z_i})^{-1}\mathbf{Z}_{z_i}^T\mathbf{K}\mathbf{W}_{z_i}$ and $\bar{\mathbf{R}}$, where $\bar{\mathbf{R}}$ is the vector of the binned population means of $\mathbf{R} = \mathbf{Y} - \mathbf{XB}$. Taking Taylor expansion on $\hat{g}(z_i)$ at $\mathbf{A}_i$ and $\bar{\mathbf{R}}$, we have

$$\hat{g}(z_i) = g(z_i) + (\hat{\mathbf{A}}_i - \mathbf{A}_i)\bar{\mathbf{R}} + \mathbf{A}_i(\hat{\bar{\mathbf{R}}} - \bar{\mathbf{R}}) .$$

Defining $\hat{\mathbf{C}} = \mathbf{K}\hat{\mathbf{W}} - \mathbf{K}\mathbf{W}$ and expanding the inverse of the matrix in $\hat{\mathbf{A}}_i$ with the expansion of the inverse of two matrices, we can show that $\hat{\mathbf{A}}_i$ is asymptotically unbiased for $\mathbf{A}_i$. Since $\hat{\bar{\mathbf{R}}}$ is a vector of binned sample means, it is asymptotically unbiased estimator of $\bar{\mathbf{R}}$. Hence, the asymptotic expectation of $\hat{g}(z_i)$ is $g(z_i)$ and the design based variance of $\hat{g}(z_i)$ is

$$V(\hat{g}(z_i)) = \mathbf{A}_i\, V(\hat{\bar{\mathbf{R}}})\mathbf{A}_i^T \tag{14}$$

where $V(\hat{\bar{\mathbf{R}}}) \approx (\mathbf{Q}\otimes\mathbf{I_m})\,COV(\hat{\bar{\mathbf{X}}},\hat{\bar{\mathbf{y}}})(\mathbf{Q}\otimes\mathbf{I_m})^T + \bar{\mathbf{X}}\,V(\hat{\mathbf{B}})\bar{\mathbf{X}}^{\mathbf{T}}$ given that $\bar{\mathbf{X}}$ is the $m \times p$ matrix of the form ($\bar{\mathbf{x}}_1^T$, …, $\bar{\mathbf{x}}_p^T$), $\hat{\bar{\mathbf{X}}}$ is the $m \times p$ matrix of the form ($\hat{\bar{\mathbf{x}}}_1^T$, …, $\hat{\bar{\mathbf{x}}}_p^T$) and **Q=** (1, -$B_1$, …, - B_p). Replacing all the variances and population parameters in equation (14), we obtain variance estimator of $V(\hat{g}(z_i))$ as

$$V(\hat{g}(z_i)) = \hat{\mathbf{A}}_i\,\hat{V}(\hat{\bar{\mathbf{R}}})\hat{\mathbf{A}}_i^T$$

where $\hat{V}(\hat{\bar{\mathbf{R}}}) \approx (\hat{\mathbf{Q}}\otimes\mathbf{I_m})\,\hat{COV}(\hat{\bar{\mathbf{X}}},\hat{\bar{\mathbf{y}}})(\hat{\mathbf{Q}}\otimes\mathbf{I_m})^T + \hat{\bar{\mathbf{X}}}\,\hat{V}(\hat{\mathbf{B}})\hat{\bar{\mathbf{X}}}^{\mathbf{T}}$.

Additionally, we establish the asymptotic normality of $\hat{g}(z_i)$, that is,

Theorem 2:

$\sqrt{n}(\hat{g}(z_i) - g(z_i))$ converges to normal with mean 0 and variance of $\mathbf{V}(\hat{g}(z_i))$ and $\mathbf{V}(\hat{g}(z_i))$ converges to $\mathbf{V}(\hat{g}(z_i))$ in probability.

4. DATA ANALYSIS

In this analysis, we illustrate semiparametric partial linear regression model with data from the Ontario Health Survey (OHS). The Ontario Health Survey was conducted with a stratified two-stage clustered design. The strata were the public health units in the province of Ontario and within each stratum neighborhoods were randomly selected as were households within each neighborhood. The purpose of this survey is to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of mortality in Ontario.

For the purpose of illustrating the partial linear model, we examine the effects of age, gender, smoking status and physical activeness on the body mass index (BMI) and the desired body mass index (DBMI). The BMI is a measure of actual weight status and the DBMI is a measure of desired weight measure. Both of the BMI and the DBMI are calculated as follows:

$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in meters})^2}$$

$$\text{DBMI} = \frac{\text{desired weight in kg}}{(\text{height in meters})^2}$$

We use age as a continuous variable and treat the other factors as discrete variables. The working model is defined as:

$$\text{BMI} = g_1(\mathbf{age}) + \mathbf{X\beta} + \boldsymbol{\varepsilon}_1$$
$$\text{DBMI} = g_2(\mathbf{age}) + \mathbf{X\beta} + \boldsymbol{\varepsilon}_2$$

where **X** is the design matrix including all the indicator variables from the factors.

Among all the explanatory variables, we focus on the continuous variable -- age. Since the BMI is not applicable to adolescents, we only pick the respondents whose ages are between 18 and 64. After deleting all the missing values and "not stated" observation, in the data file there are a total of 21968 observations. Since there are only 46 distinct points in the age variable, we bin the data set according to age. The bin size is set to be unity such that there are 46 bins with midpoints being 18,19,…, 64.

Table 1 lists all the survey estimates of the linear coefficients of the first model. On comparing BMI by gender, we found that male BMI is higher. Using former smoker as the base category, the coefficients of the smoking status are all negative and significant, which suggests that former smokers tend to be heavier than people with other types of smoking status.

Table 1: Estimates of the Linear Coefficients

Factors	$\hat{\mathbf{B}}$	**SE(** $\hat{\mathbf{B}}$ **)**	***t*-value**
Gender	1.45	0.052	27.90
Never Smoked	-1.45	0.065	-22.27
Occasional Smoker	-1.72	0.12	-14.41
Daily Smoker	-1.48	0.072	-20.63
Moderate Active	0.66	0.095	6.96
Inactive	1.43	0.078	18.45

In **Figure 1** and **Figure 2**, the estimated functions of age, $\hat{g}_1(\mathbf{age})$ and $\hat{g}_2(\mathbf{age})$, and their confidence intervals are plotted versus different ages. It is found that, in both cases, the BMI and the DBMI are increasing nonlinear functions of age. A comparison of $\hat{g}_1(\mathbf{age})$ with $\hat{g}_2(\mathbf{age})$ is shown in **Figure 3**. It is found on average that for every individual who is either active or moderate active, the DBMI is lower than the BMI.

Figure 1: Estimated Age Trend in BMI with Confidence Intervals

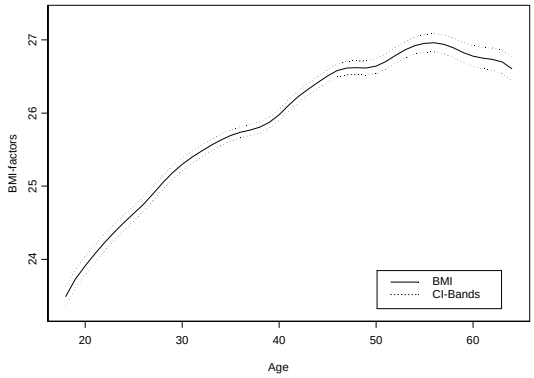


Figure 2: Estimated Age Trend in DBMI with Confidence intervals

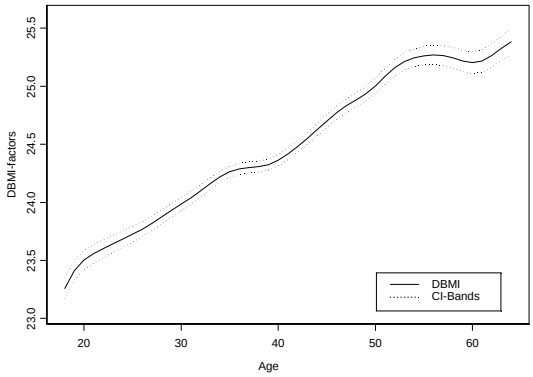
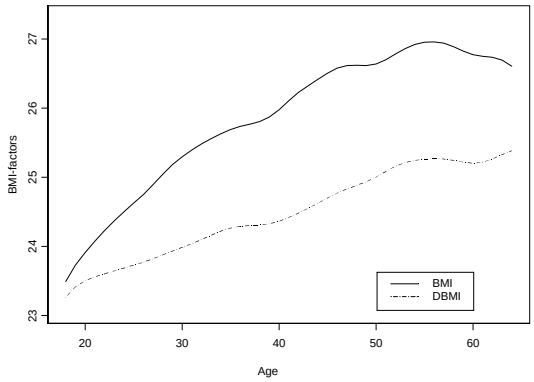


Figure 3: Estimated Age Trend in BMI and DBMI



5. CONCLUSION

With the assistance of a partial linear model, we extend semi-parametric regression techniques to complex survey data. Asymptotic properties of the survey estimators are developed. Computation of the variance estimates of both linear coefficients and the regression function rely on variance of survey total and means. Provided that we obtain the required variance estimates, we can apply this method using standard statistical packages. In the partial linear working model, we assume that there is no interaction between the parametric component and the nonparametric component. This assumption can be relaxed in such a way that nonparametric component appears linearly in the interaction term. When estimating conditional expectation on the nonparametric components for indicator discrete random variables, we propose to use generalized linear or additive models to conduct the estimation.

REFERENCES

Bellhouse, D.R. and Stafford, J.E. (1999), Density estimation from complex survey, *Statistica Sinica*, 9, pp. 407-424.

Bellhouse D. R. and Stafford, J. E. (2001), Local polynomial regression in complex survey. *Survey methodology*, 27(2), pp. 197 - 203.

Binder, D. A.(1983), On the variance of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, pp. 279-292.

Jones, M. C. (1989), Discretized and interpolated kernel density estimates, *Journal of the American Statistical Association*, 84, pp.733-741.

Konijn, H. S. (1962), Regression analysis in the sample surveys, *Journal of the American Statistical Association*, 57, pp. 590-606.

Krewski, D. and J. Rao (1981), Inference from stratified samples: Properties of the linearization, jackknife and balance repeated replication methods, *The Annals of Statistics*, 9(5), pp. 1010-1019.

Ontario Ministry of Health (1996), *Ontario Health Survey: User's Guide, Volumes I and II*, Queen's Printer for Ontario.

Robinson, P. M. (1988), Root-N-consistent semiparametric regression, *Econometrica*, 56, pp. 931-954.

Shao, J. (1996), Resampling methods in sample survey, *Statistics*, 27, pp. 203-254.

Speckman, P. (1988), Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society, Series B (Methodological)*, 50(3), pp. 413-436.