

SEMI-PARAMETRIC ANALYSIS OF THE COVARIANCE FOR COMPLEX SURVEY DATA

Zilin Wang and David R. Bellhouse¹

Abstract

The aim of this work is to develop an estimation method for a semi-parametric regression model for large-scale surveys. In this semi-parametric regression model, the explanatory variables are represented separately in two parts: the nonparametric part and the parametric linear part. We estimate both the function of the nonparametric part of the model and the parameters that are included in the parametric part of the model. The estimating method for this model combines the local polynomial regression in complex surveys by Bellhouse and Stafford (1999) and the classical least square estimate. Moments of the related estimates have been derived. An empirical illustration is carried out using the 1990 Ontario Health survey.

KEY WORDS: Binning; Ontario health survey; Regression; Sampling; Smoothing.

Résumé

Le but de ce travail est de développer une méthode d'estimation pour un modèle de régression semi-paramétrique pour un sondage de grande envergure. Dans ce modèle de régression semi-paramétrique, les variables explicatives sont représentées séparément en deux parties: la partie non paramétrique et la partie linéaire paramétrique. Nous estimons la fonction de la partie non paramétrique du modèle et aussi les paramètres inclus dans la partie paramétrique du modèle. La méthode d'estimation pour ce modèle regroupe la régression polynomiale locale pour des sondages complexes donnée par Bellhouse et Stafford(1999) et l'estimation des moindres carrés classiques. Les moments des estimateurs ont été dérivés. Une illustration empirique est effectuée en utilisant l'Enquête sur la santé en Ontario de 1990.

MOTS CLÉS: Binning; échantillonnage; Enquête sur la santé en Ontario; lissage; régression.

1. INTRODUCTION

As in all areas of statistics field, regression analysis plays an important role in the survey sampling. A typical regression model is of the form:

$$E(Y|X) = g(X, \beta)$$

where functional form of the $g(\cdot, \cdot)$ is known and β is an unknown p -dimensional parameter. In regression analysis theory, correct specification of the model is a major problem. Over the years, questions related to the specification of the regression model have attracted the attention of many researchers. There exist various alternative methods to the conventional parametric (linear and nonlinear) regression models. Among all the existing methods, nonparametric regression models are very commonly used. A nonparametric regression model can be estimated by a smoother. However, in a p -dimensional multiple regression model, despite the

difficulty of choosing the right window size, a more serious problem that relates to all smoothing methods is the "curse of dimensionality". Neighborhoods with a fixed number of points become less local as the dimensions increase. Several multivariate nonparametric regression techniques have been devised. As suggested by the word "nonparametric" in a nonparametric regression, we do not make any assumption on the functional form of the model. Instead of estimating some parameters of the model, we estimate the conditional expectation of the response Y on the predictors, $E(Y|X)$ or the function $g(\cdot, \cdot)$.

It is well-known that there are multiple observations at many of the distinct values for a large-scale survey data. It has been shown by Bellhouse and Stafford (1999) that the nonparametric approach of local

¹ Zilin Wang and David R. Bellhouse, Department of Statistics and Actuarial Science, University of Western Ontario, London, Ontario, Canada.

polynomial regression model can be applied to a large-scale survey by kernel smoothing. The method introduced by Bellhouse and Stafford (1999) enables us to estimate the conditional expectation in a regression model for complex surveys. However, as in all of nonparametric estimation, "curse of dimensionality" is still a possible barrier to local polynomial methods being used for high dimensional regression models.

In this paper, a so-called "partial linear semiparametric regression model" will be introduced for complex surveys. In this semiparametric regression model, the explanatory variables are represented separately in two parts: the nonparametric part and the parametric linear part. We are interested in estimating both the functional form of the nonparametric part of the model and the parameters that included in the parametric part of the model. The estimation method for this model combines the local polynomial regression in complex survey by Bellhouse and Stafford (1999) and the classical least squares estimate. 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 the nonparametric regression estimation because of low effective sample sizes. It is a very natural to include the discrete explanatory variables in the linear part of the model. The estimation method of the partial linear semiparametric model and the related theoretical properties in the case of independently and identically distributed random variables was first introduced by Robinson (1988). It is of interest to adopt this estimation method to complex survey data.

The paper is organized as follows. In section 2, we introduce the estimation procedures. In section 3, an empirical illustration of the estimation method is carried out using the 1990 Ontario Health Survey in Section 4.

2. ESTIMATION OF THE PARTIAL LINEAR SEMIPARAMETRIC MODEL

Consider a model-based partial linear regression model in the complex survey defined as:

$$\mathbf{y} = g(\mathbf{z}) + \mathbf{X}\beta + \varepsilon \quad (1)$$

where \mathbf{y} is the response variable and both \mathbf{X} and \mathbf{z} are the explanatory variables. With the finite population size N , \mathbf{X} is a $N \times p$ matrix, \mathbf{z} is measured on a

continuous scale and $g(\cdot)$ is an arbitrary univariate function of \mathbf{z} . Both the functional form of $g(\cdot)$ and parameters β are unknown. Additionally, we assume that $E(\varepsilon|\mathbf{z}, \mathbf{X}) = \mathbf{0}$ and that there is no interaction between \mathbf{X} and \mathbf{z} .

The problem in estimating β in the partial linear model as stated in (1) is that there is a function of unknown form, $g(\mathbf{z})$. If it were possible to find a means of removing this function, then we would be left with a linear regression model which could then be estimated by using least squares. On taking the expectation of \mathbf{y} in (1) on conditional \mathbf{z} , we obtain

$$E(\mathbf{y}|\mathbf{z}) = E(\mathbf{X}|\mathbf{z})\beta + g(\mathbf{z}) \quad (2)$$

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

$$\mathbf{y} - E(\mathbf{y}|\mathbf{z}) = (\mathbf{X} - E(\mathbf{X}|\mathbf{z}))\beta + \varepsilon \quad (3)$$

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

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (4)$$

Now one obvious approach is to estimate β by the method of least squares. Unfortunately, since $E(\mathbf{y}|\mathbf{z})$ and $E(\mathbf{X}|\mathbf{z})$ are unknown, least squares estimation of β is not feasible. Consequently, we carry out the estimation of β in two steps. In step 1, we estimate the conditional expectations appearing in the (3) using the local polynomial regression techniques for complex survey developed by Bellhouse and Stafford (1999). In step 2, we replace $E(\mathbf{y}|\mathbf{z})$ and $E(\mathbf{X}|\mathbf{z})$ in (3) with their estimates obtained in step 1 and estimate β with the method of least squares.

2.1. Estimation of the Parametric Part of Model

In order to accomplish the estimation of the conditional expectations, we bin the observed data according to \mathbf{z} . Suppose that \mathbf{z} has m distinct values in the finite population of size N . Let z_i denote the i^{th} distinct value or the i^{th} bin and assume that the values of \mathbf{z} 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 \mathbf{y} at distinct values of \mathbf{z} be $\bar{\mathbf{y}} = (\bar{y}_1, \dots, \bar{y}_m)$ and the vector of finite population means for the j^{th} independent variable \mathbf{x}_j for $j = 1, \dots, p$ be $\bar{\mathbf{x}}_j = (\bar{x}_{1j}, \dots, \bar{x}_{mj})$. \hat{y}_i , \hat{x}_{ij} and \hat{p}_i are the survey estimates of \bar{y}_i , \bar{x}_{ij} and \bar{p}_i , respectively. We estimate the conditional expectation at each distinct point of \mathbf{z} . At any one of z_i , for all $i = 1, \dots, m$, let

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

$\mathbf{KW}_{z_i} = \frac{1}{h} \text{diag}(p_1 K(\frac{z_1 - z_i}{h}), \dots, p_m K(\frac{z_m - z_i}{h}))$ where

$K(\cdot)$ is a kernel function and h is the bandwidth. Thus, the finite population estimates of the expectation of \mathbf{X} and \mathbf{y} conditional on $z = z_i$ are:

$$E(\mathbf{X} | z = z_i) = \mathbf{m}_X(z_i) = \mathbf{A}_i \bar{\mathbf{X}} \quad (5)$$

$$E(\mathbf{y} | z = z_i) = m_y(z_i) = \mathbf{A}_i \bar{\mathbf{y}} \quad (6)$$

where $\mathbf{A}_i \equiv \mathbf{e}^T (\mathbf{Z}_{z_i}^T \mathbf{KW}_{z_i} \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^T \mathbf{KW}_{z_i}$. The vector \mathbf{e} is the $m \times 1$ vector of the form $(1, 0, \dots, 0)^T$, $\bar{\mathbf{y}}$ is the $m \times 1$ vector of the form $(\bar{y}_1, \dots, \bar{y}_m)^T$, and $\bar{\mathbf{X}}$ is the $m \times p$ matrix of the form $(\bar{\mathbf{x}}_1^T, \dots, \bar{\mathbf{x}}_p^T)$. The sampling

estimate of the kernel weight matrix, $\hat{\mathbf{KW}}_{z_i}$, can be obtained by replacing p_i in the kernel weight matrix, \mathbf{KW}_{z_i} , with \hat{p}_i for $i = 1, \dots, m$. If we then replace $\bar{\mathbf{X}}$, $\bar{\mathbf{y}}$ with $\hat{\mathbf{X}}$, $\hat{\mathbf{y}}$ in (5) and (6), respectively, we have the analogous sampling estimates,

$$\hat{E}(\mathbf{X} | z = z_i) = \hat{\mathbf{m}}_X(z_i) = \hat{\mathbf{A}}_i \hat{\mathbf{X}} \quad (7)$$

$$\hat{E}(\mathbf{y} | z = z_i) = \hat{m}_y(z_i) = \hat{\mathbf{A}}_i \hat{\mathbf{y}} \quad (8)$$

where $\hat{\mathbf{A}}_i \equiv \mathbf{e}^T (\mathbf{Z}_{z_i}^T \hat{\mathbf{KW}}_{z_i} \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^T \hat{\mathbf{KW}}_{z_i}$. The vector $\hat{\mathbf{y}}$ is the $m \times 1$ vector of the form $(\hat{y}_1, \dots, \hat{y}_m)^T$, and $\hat{\mathbf{X}}$ is the $m \times p$ matrix of the form $(\hat{\mathbf{x}}_1^T, \dots, \hat{\mathbf{x}}_p^T)$.

Using the finite population estimates of the conditional expectations, we reconstruct our data. The reason why we reconstruct the data is to take advantage the sampling design characteristic of the original data. Let N_i be number of observations that fall in the i^{th} bin and $\sum_{i=1}^m N_i = N$. $\mathbf{M}_X(z)$ is a $N \times p$ matrix consisting of all the estimated conditional expectations of \mathbf{X} and $\mathbf{M}_y(z)$ is a $N \times 1$ vector such that $m_y(z_i)$ is repeated for N_i times in the i^{th} bin, that is,

$$\mathbf{M}_X(z) = \begin{pmatrix} \begin{pmatrix} m_{x_1}(z_1) & m_{x_2}(z_1) & \cdots & m_{x_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ m_{x_1}(z_m) & m_{x_2}(z_m) & \cdots & m_{x_p}(z_m) \end{pmatrix}_{N_i \times p} \\ \vdots \\ \begin{pmatrix} m_{x_1}(z_m) & m_{x_2}(z_m) & \cdots & m_{x_p}(z_m) \\ \vdots & \vdots & \vdots & \vdots \\ m_{x_1}(z_m) & m_{x_2}(z_m) & \cdots & m_{x_p}(z_m) \end{pmatrix}_{N_m \times p} \end{pmatrix}$$

$$\mathbf{M}_y(z) = \begin{pmatrix} \begin{pmatrix} m_y(z_1) \\ \vdots \\ m_y(z_1) \end{pmatrix}_{N_i \times 1} \\ \vdots \\ \begin{pmatrix} m_y(z_m) \\ \vdots \\ m_y(z_m) \end{pmatrix}_{N_m \times 1} \end{pmatrix}$$

Using $\mathbf{M}_X(z)$ and $\mathbf{M}_y(z)$, we can transform our data as: $\mathbf{Y} \equiv \mathbf{y} - \mathbf{M}_y(z)$ and $\mathbf{X} \equiv \mathbf{X} - \mathbf{M}_X(z)$. The multiple regression least square estimation without constant term for the finite population is:

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

If we have a set of sampling data $(\mathbf{X}^s, \mathbf{y}^s)$ with sample size n and n_i observations within each bin such that $\sum_{i=1}^m n_i = n$, we can construct $\hat{\mathbf{M}}_X(z)_{n \times p}$ and $\hat{\mathbf{M}}_y(z)_{n \times 1}$ with the same way as we construct $\mathbf{M}_X(z)$ and $\mathbf{M}_y(z)$. We use sampling estimates $\hat{m}_{x_j}(z_i)$ and $\hat{m}_y(z_i)$ that are shown in (7) and (8) instead of $m_{x_j}(z_i)$ and $m_y(z_i)$. Defining $\hat{\mathbf{Y}} = \mathbf{y}^s - \hat{\mathbf{M}}_y(z)$ and $\hat{\mathbf{X}} = \mathbf{X}^s - \hat{\mathbf{M}}_X(z)$, we have the corresponding least square estimators in the context of complex survey:

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

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

2.2 Estimation of the Nonparametric Part of the Model

Once we obtain the finite population parameter \mathbf{B} , we can estimate the population estimate $g(\cdot)$ by using the following model,

$$\mathbf{y} - \mathbf{XB} = \mathbf{g}(z) + \mathbf{v} \quad (11)$$

By letting $\mathbf{R} = \mathbf{y} - \mathbf{XB}$ and $\bar{\mathbf{R}}$ be the population mean vector of \mathbf{R} for each bin. By applying the local polynomial technique again, we have

$$g(z_i) = \mathbf{A}_i \bar{\mathbf{R}}$$

Using the sampling estimates $\hat{\mathbf{B}}$, we have $\bar{\mathbf{R}} = \mathbf{y}^s - \mathbf{X}^s \hat{\mathbf{B}}$. The sampling estimate of $g(z_i)$ is,

$$\hat{g}(z_i) = \hat{A}_i \hat{R}$$

where \hat{R} is the survey estimates of \bar{R} .

3. 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 and physical activity 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 following:

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

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

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

$$BMI = g_1(\text{age}) + \mathbf{X}B + \varepsilon_1$$

$$DBMI = g_2(\text{age}) + \mathbf{X}B + \varepsilon_2$$

where \mathbf{X} is the design matrix including all the indicator variables.

Among all the explanatory variables, we focus on the continuous variable -- age. Since the BMI is not applicable to adolescents, we only pick the observations between age 18 to age 64. After deleting all the missing values and "not stated" observation, it leaves us a total of 21968 observations. However, 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 unit, thus, there are 46 bins with midpoints being 18,19,..., 64. In **Figure 1** and **Figure 2**, the estimated functions of age, $\hat{g}_1(\text{age})$ and $\hat{g}_2(\text{age})$, and their confidence bands are plotted versus different ages. It is found that, in both cases, the BMI and the DBMI are increasing functions of age. A comparison of the $\hat{g}_1(\text{age})$ with the $\hat{g}_2(\text{age})$ is shown in **Figure 3**. It is found on average that for

Figure 1: Estimated Age Trend in BMI with Confidence Bands

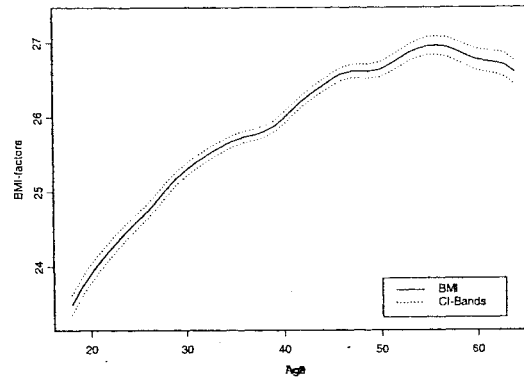


Figure 2: Estimated Age Trend in DBMI with Confidence Bands

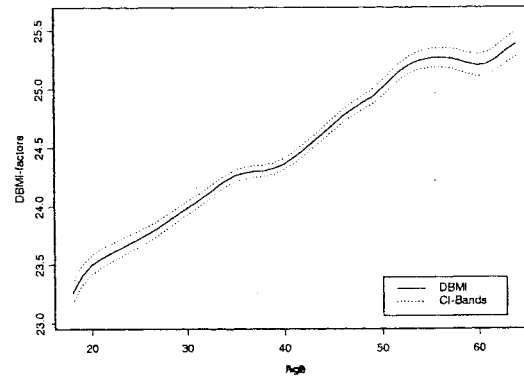
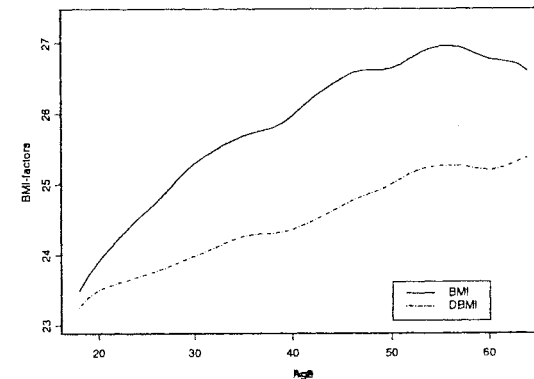


Figure 3: Estimated Age Trend in BMI and DBMI



every individual who is either active or moderate active, the DBMI is lower than the BMI.

REFERENCES

- Bellhouse, D.R. and Stafford, J.E. (1999). Density estimation from complex survey. *Statistica Sinica*. 9: 407--424.
- Bellhouse D. R. and Stafford, J. E. (1999). Local polynomial regression in complex survey. *University of Western Ontario Technical Report*.
- Binder, D. A. (1983). On the variance of asymptotically normal estimators from complex

- surveys. *International Statistical Review*. **51**: 279-292.
- Honijn, H. S. (1962). Regression analysis in the sample surveys. *Journal of the American Statistical Association*. **57**: 590-606.
- Jones, M. C. (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, **84**: 733-741.
- 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**: 931-954.