Vol. 35 (2015)
No. 5

数 学 杂 志
J. of Math. (PRC)

STATISTICAL ESTIMATION IN NONLINEAR SEMIPARAMETRIC EV MODELS WITH VALIDATION DATA

XIAO Yan-ting [1,2], TIAN Zheng [1], SUN Jin [2]

(*1.Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an 710129, China*)

(*2.Department of Applied Mathematics, Xi'an University of Technology, Xi'an 710054, China*)

Abstract: In this paper, nonlinear semiparametric error-in-variables(EV) models are considered with validation data. Without specifying any error structure equation, two estimators for the parameter in the nonlinear function are proposed based on the least square method and the kernel smoothing technique. The obtained estimators are proved to be asymptotically normal. A simulation study is conducted to show the proposed estimation methods are valid in finite sample.

Keywords: nonlinear semiparametric EV model; validation data; asymptotic normality

2010 MR Subject Classification: 62G05

Document code: A **Article ID:** 0255-7797(2015)05-1075-11

1 Introduction

Consider the nonlinear semiparametric model

$$Y = g(X, \beta) + m(T) + e, \tag{1}$$

where Y is the scalar response variable, X is a p-dimensional covariate and T is a univariate random variable, $g(x, \beta)$ is a pre-specified function in which β is an unknown parameter vector in R^d and $m(.)$ is an unknown smooth function. The model error e are independent and identically distributed with zero mean. Obviously, model (1) is reduced to be a partially linear model if let $g(X, \beta) = X^T\beta$.

Model (1) is a very extensive semiparametric model which was widely studied in many fields, such as econometric, biology, and environmental science. Li and Nie [1] proposed an estimation procedure for parameter β through a nonlinear mixed-effects approach. Furthermore, Li and Nie [2] analyzed a real data in ecology with this model and proposed two estimation procedures by profile nonlinear least squares and linear approximation approach. Huang and Chen [3] obtained the spline profile least square estimator of parameter β when

* **Received date:** 2014-03-27 **Accepted date:** 2014-12-18

Foundation item: Supported by National Natural Science Foundation of China(61303223; 91120014).

Biography: Xiao Yanting(1981–), female, born at Xi'an, Shaanxi, lecturer, major in nonparametric statistics.

the baseline function $m(.)$ was approximated by some graduating functions. Later, Song et al. [4] provided a sieve least square method when the nonlinear function $g(.,.)$ has some special form. Recently, Xiao et al. [5] applied empirical likelihood approach to this model and compared with the normal approximation method in terms of confidence region of parameter β.

In practice, some variables of our interest are difficult or expensive to be measured exactly and then are usually replaced by some surrogate observations. The semiparametric errors-in-variables (EV) model has frequently been applied to many fields and has received much attention in the literature. The initial assumption is that the variable error is additive. [6]–[9] applied the empirical likelihood method to partially linear models and varying-coefficient partially linear models with additive error assumption. However, the additive error assumption is usually not appropriate in real situation. The realistic case is that the relationship between the surrogate variables and the true variables is rather complicated and may be that no error model structure is assumed. In this case, one solution is employing the help of validation data to capture the underlying relation between the true variables and surrogate variables.

When the error existed in the covariables, some statistical inference based on validation data were developed. Wang [10] used this method to partially linear error-in-variable model. Wang and Rao [11] and Stute et al. [12] developed empirical likelihood approach to linear models and nonlinear models with errors-in-covariables, respectively. Recently, Wang and Zhang [13] and Du et al. [14] applied statistical inference to varying coefficient models and nonparametric regression function with validation sampling. Later, Fang and Hu [15] considered the nonlinear model with the help of validation data when the error is in the response. For nonlinear semiparametric models, Xue [16] constructed empirical log-likelihood ratio statistics for the unknown parameter with the help of validation data. Furthermore, Liu [17] considered nonlinear semiparametric models with missing response variable and error-in-covariables.

In this paper, we consider model (1) with explanatory variable X measured with error and both Y and T measured exactly. Instead of the true variable X, the surrogate variable $\tilde{X}$ is observed. The relationship between X and $\tilde{X}$ is not additive, which can be evaluated by regression of X on $\tilde{X}$. This assumption has been used in other statistical models, such as in linear models [11] and varying coefficient models [13]. We define two estimators for the parameter in nonlinear function by considering the two cases where the response variable Y is available or not in the validation sample. Asymptotic results for the two estimators are derived, showing that the two proposed estimators are asymptotically normal.

The rest of this paper is organized as follows: we describe the estimation procedures based on the least square method and kernel method in Section 2. In Section 3, the asymptotic normality of the proposed estimators is proved. Some simulation studies are conducted in Section 4 to evaluate the finite sample properties of the proposed estimators. Finally, Section 5 concludes the paper.

2 Estimation

Suppose that $\tilde{X}$ is a p-dimensional surrogate variable for X. Assume that we have a primary data set containing N independent and identically distributed observations of $\{(Y_j, \tilde{X}_j, T_j)_{j=n+1}^{n+N}\}$ and a validation data set containing n independent and identically distributed observations of $\{(X_i, \tilde{X}_i, T_i)_{i=1}^{n}\}$ or $\{(Y_i, X_i, \tilde{X}_i, T_i)_{i=1}^{n}\}$. It is also assumed that the two observation subsets are independent.

Denote $Z = (\tilde{X}, T)$ and $G(z, \beta) = E[g(X, \beta)|Z = z]$. Then, model (1) can be rewritten as

$$Y = G(Z, \beta) + m(T) + \varepsilon, \tag{2}$$

where $\varepsilon = e + g(X, \beta) - G(Z, \beta)$.

Clearly, model (2) is a standard partially nonlinear model if $G(.,.)$ is a known function. Unfortunately, $G(.,.)$ is usually unknown in practice. To solve this difficulty, we estimate $G(.,.)$ consistently by the kernel method with validation data as following procedure.

Let

$$\hat{R}_n(z, \beta) = \frac{1}{nh_{1,n}} \sum_{i=1}^{n} g(X_i, \beta) K_1\big(\frac{Z_i - z}{h_{1,n}}\big), \qquad \hat{f}_n(z) = \frac{1}{nh_{1,n}} \sum_{i=1}^{n} K_1(\frac{Z_i - z}{h_{1,n}}),$$

where $K_1(.)$ is a kernel function and $h_{1,n}$ is a bandwidth.

Then, $G(z, \beta)$ can be estimated by $\frac{\hat{R}_n(z,\beta)}{\hat{f}_n(z)}$. Notice that the small value of $\hat{f}_n(z)$ as the denominator in this estimator, so we can improve this estimator in practice to avoid technical difficulties. Let $\hat{f}_{nb}(z) = \max(\hat{f}_n(z), b_n)$, where b_n is a positive constant sequence that decrease to zero as n increase to infinity. Then, the estimator of $G(z, \beta)$ with truncation version, say $\hat{G}(z, \beta)$, is given by

$$\hat{G}(z, \beta) = \frac{\hat{R}_n(z, \beta)}{\hat{f}_{nb}(z)}. \tag{3}$$

Define $G^{(1)}(z, \beta) = \frac{\partial}{\partial \beta} G(z, \beta) = E[g^{(1)}(X, \beta)|Z = z]$ and $g^{(1)}(X, \beta) = \frac{\partial}{\partial \beta} g(X, \beta) = \big(\frac{\partial}{\partial \beta_1} g(X, \beta), \cdots, \frac{\partial}{\partial \beta_d} g(X, \beta)\big)^T$. Then, the estimator of $G^{(1)}(z, \beta)$, denoted by $\hat{G}^{(1)}(z, \beta)$, can also be obtained by the kernel method.

Let

$$\hat{R}_n^{(1)}(z, \beta) = \frac{1}{nh_{1,n}} \sum_{i=1}^{n} g^{(1)}(X_i, \beta) K_1(\frac{Z_i - z}{h_{1,n}}),$$

then, we have

$$\hat{G}^{(1)}(z, \beta) = \frac{\hat{R}_n^{(1)}(z, \beta)}{\hat{f}_{nb}(z)}. \tag{4}$$

Using $\hat{G}(z, \beta)$ to replace $G(z, \beta)$ in model (2) and assuming β is known, $m(t)$ is estimated by

$$\hat{m}(t, \beta) = \sum_{j=n+1}^{n+N} W_{Nj}(t)[Y_j - \hat{G}(Z_j, \beta)], \tag{5}$$

where $W_{Nj}(t)=\frac{K_2(\frac{T_j-t}{h_{2,N}})}{\sum\limits_{i=n+1}^{n+N}K_2(\frac{T_i-t}{h_{2,N}})}$ with $K_2(.)$ is a kernel function and $h_{2,N}$ is a bandwidth.

Similar to $\hat{m}(t,\beta)$ defined in (5), the estimator of $E[G^{(1)}(Z,\beta)|T=t]$, denoted by $\hat{h}(t,\beta)$, can be estimated by the kernel method, which is defined as

$$\hat{h}(t,\beta)=\sum_{j=n+1}^{n+N}W_{Nj}(t)\hat{G}^{(1)}(Z_j,\beta). \tag{6}$$

Then, the estimator of β is defined to be the one which minimizes $\hat{S}_N(\beta)$ given by

$$\hat{S}_N(\beta)=\frac{1}{N}\sum_{j=n+1}^{n+N}(Y_j-\hat{G}(Z_j,\beta)-\hat{m}(T_j,\beta))^2. \tag{7}$$

Thus, the estimator of β, say $\hat{\beta}_N$, solves the equation

$$\frac{1}{N}\sum_{j=n+1}^{n+N}(Y_j-\hat{G}(Z_j,\beta)-\hat{m}(T_j,\beta))(\hat{G}^{(1)}(Z_j,\beta)-\hat{h}(T_j,\beta))=0. \tag{8}$$

Notice that, if we ignore the missing response variable in Liu [17], the estimator $\hat{\beta}$ will reduce to be the $\hat{\beta}_N$ in this paper. In practice, the response variable Y may be fully observed, that is to say Y can also be measured in the validation data set. In this case, considering the validation data $\{(Y_i,X_i,\tilde{X}_i,T_i)_{i=1}^n\}$, an alternative estimator of β, say $\hat{\beta}_{n,N}$, can be obtained by following procedures.

Let

$$\tilde{m}(t,\beta)=\sum_{i=1}^{n}\tilde{W}_{ni}(t)[Y_i-g(X_i,\beta)], \tag{9}$$

where $\tilde{W}_{ni}(t)=\frac{K_3(\frac{T_i-t}{h_{3,n}})}{\sum\limits_{j=1}^{n}K_3(\frac{T_j-t}{h_{3,n}})}$ with $K_3(.)$ is a kernel function and $h_{3,n}$ is a bandwidth.

Similar to (9), the estimator of $E[g^{(1)}(X,\beta)|T=t]$, denoted by $\tilde{h}(t,\beta)$, is defined as

$$\tilde{h}(t,\beta)=\sum_{i=1}^{n}\tilde{W}_{ni}(t)g^{(1)}(X_i,\beta). \tag{10}$$

Then, $\hat{\beta}_{n,N}$ can be obtained by minimizing the sum of least squares

$$\begin{aligned}\hat{S}_{n,N}(\beta)=&\frac{1}{n+N}\bigg\{\sum_{j=n+1}^{n+N}(Y_j-\hat{G}(Z_j,\beta)-\hat{m}(T_j,\beta))^2\\&+\sum_{i=1}^{n}(Y_i-g(X_i,\beta)-\tilde{m}(T_i,\beta))^2\bigg\}.\end{aligned} \tag{11}$$

Thus, $\hat{\beta}_{n,N}$ solves the equation

$$\begin{aligned}&\frac{1}{n+N}\Bigg\{\sum_{j=n+1}^{n+N}(Y_j-\hat{G}(Z_j,\beta)-\hat{m}(T_j,\beta))(\hat{G}^{(1)}(Z_j,\beta)-\hat{h}(T_j,\beta))\\&+\sum_{i=1}^{n}(Y_i-g(X_i,\beta)-\tilde{m}(T_i,\beta))(g^{(1)}(X_i,\beta)-\tilde{h}(T_i,\beta))\Bigg\}=0.\end{aligned}\tag{12}$$

Finally, using estimator $\hat{\beta}_N$ or $\hat{\beta}_{n,N}$, we can define the estimator of $m(.)$ as following

$$\hat{m}_N(t)=\sum_{j=n+1}^{n+N}W_{Nj}(t)(Y_j-\hat{G}(Z_j,\hat{\beta}_N)),\tag{13}$$

$$\hat{m}_{n,N}(t)=\sum_{j=n+1}^{n+N}W_{Nj}(t)(Y_j-\hat{G}(Z_j,\hat{\beta}_{n,N})).\tag{14}$$

3 Asymptotic Property

To state our results, we introduce the following assumptions:

(A1) $m(t)$ has two bounded and continuous derivatives on (0,1).

(A2) T has density function $r(t)$ on $[0,1]$, and $0<\inf_{0\le t\le 1} r(t)<\sup_{0\le t\le 1} r(t)<\infty$.

(A3) $\sup_z E[e^2|Z=z]<\infty$, $\sup_z E[g^2(X,\beta)|Z=z]<\infty$, $\sup_z E[g_s^{(1)}(X,\beta)^2|Z=z]<\infty$, $s=1,2,\cdots,d$.

(A4) For some $k>p$, $G(z,\beta)\in\Re^k$, and $G_s^{(1)}(z,\beta)\in\Re^k$.

(A5) The density of Z, say $f_Z(z)$, has bounded partial derivative of order one and satisfies $NP(f_z(z)<\eta_N)\to 0$ for some positive constant sequence $\eta_N>0$ tending to zero.

(A6) The kernel function $K_1(.)$ is a $d+1$-dimensional, continuous and symmetric probability density function with bounded support. Both $K_2(.)$ and $K_3(.)$ are symmetric and bounded probability density function with finite support.

(A7) $nh_{1,n}^{2p}b_n^4\to\infty$, $nh_{1,n}^{2k}b_n^{-2}\to 0$ $(k>p)$, $Nh_{2,N}\to\infty$ and $Nh_{2,N}^4\to 0$, $nh_{3,n}\to\infty$ and $nh_{3,n}^4\to 0$.

(A8) Both $\Sigma_1(\beta)$ and $\Sigma_3(\beta)$ are positive definite matrixes which defined in Theorem 1 and Theorem 2.

(A9) $\frac{N}{n}\to\lambda$, where λ is a nonnegative constant.

Remark 1 (A1), (A2), (A3), (A8) are standard assumptions in partially nonlinear regression models. (A4) and (A5) are common assumptions in measurement error data with validation sample. Assumptions (A6), (A7), (A9) are usual used in kernel function and bandwidths assumptions.

For the estimator $\hat{\beta}_N$, asymptotic normality is given by the following theorem.

Theorem 1 Under assumptions A1–A9, we have

$$\sqrt{N}(\hat{\beta}_N-\beta)\xrightarrow{d}N(0,\Sigma_1^{-1}(\beta)[V_0(\beta)+\lambda V_1(\beta)]\Sigma_1^{-1}(\beta))$$

where $\xrightarrow{d}$ denotes the convergence in distribution, $\Sigma_1(\beta)=E[U(Z,\beta)U^T(Z,\beta)]$ with

$$\begin{aligned}&U(Z,\beta)=G^{(1)}(Z,\beta)-E[G^{(1)}(Z,\beta)|T],\\&V_0(\beta)=E\{[Y-G(Z,\beta)-m(T)]^2U(Z,\beta)U^T(Z,\beta)\},\\&V_1(\beta)=E\{[G(Z,\beta)-g(X,\beta)]^2U(Z,\beta)U^T(Z,\beta)\}.\end{aligned}$$

Proof The proof of Theorem 1 is similar to Theorem 2.3 in Xue [16], so we omit it.

Remark 2 The first term in the asymptotic covariance is the contribution of the primary data in the sample by modeling (2), the partially nonlinear regression relationship between Y, and Z, T. The second term represents the extra cost due to the estimation of unknown mean $g(X,\beta)$ given Z using the validation data. If $\lambda=0$, the second term in the asymptotic covariance will disappears, and the asymptotic covariance is the same as that in Li and Nie [2].

For the estimator $\hat{\beta}_{n,N}$, we give the following theorem.

Theorem 2 Under assumptions A1–A9, we have

$$\sqrt{N+n}(\hat{\beta}_{n,N}-\beta)\xrightarrow{d}N\left(0,\Sigma_3^{-1}(\beta)V(\beta)\Sigma_3^{-1}(\beta)\right),$$

where $\Sigma_3(\beta)=\frac{\lambda}{1+\lambda}\Sigma_1(\beta)+\frac{1}{1+\lambda}\Sigma_2(\beta)$ with

$$\begin{aligned}&\Sigma_2(\beta)=E[H(X,\beta)H^T(X,\beta)],\\&H(X,\beta)=g^{(1)}(X,\beta)-E[g^{(1)}(X,\beta)|T],\\&V(\beta)=\frac{\lambda}{1+\lambda}(V_0(\beta)+\lambda V_1(\beta))+\frac{1}{1+\lambda}V_2(\beta),\\&V_2(\beta)=E[(Y-g(X,\beta)-m(T))^2H(X,\beta)H^T(X,\beta)].\end{aligned}$$

Proof To facilitate the presentation, we give the notations as $A^{\otimes 2}=AA^T$ for a vector or matrix A. Define the left side of (12) is $K(\beta)$, that is

$$\begin{aligned}K(\beta)=&\frac{1}{n+N}\Bigg\{\sum_{j=n+1}^{n+N}(Y_j-\hat{G}(Z_j,\beta)-\hat{m}(T_j,\beta))(\hat{G}^{(1)}(Z_j,\beta)-\hat{h}(T_j,\beta))\\&+\sum_{i=1}^{n}(Y_i-g(X_i,\beta)-\tilde{m}(T_i,\beta))(g^{(1)}(X_i,\beta)-\tilde{h}(T_i,\beta))\Bigg\}\\:=&\frac{1}{n+N}[A(\beta)+B(\beta)].\end{aligned}\tag{15}$$

By the motivation of (12), we have $K(\hat{\beta}_{n,N})=0$. Using Taylor expression to $K(\beta)$ at $\hat{\beta}_{n,N}$, we get that

$$\hat{\beta}_{n,N}-\beta=C_{n,N}^{-1}(\beta^*)(K(\beta))+O_p(N^{-\frac{1}{2}}),\tag{16}$$

where $C_{n,N}(\beta)=\frac{1}{n+N}\big[\sum\limits_{j=n+1}^{n+N}(\hat{G}^{(1)}(Z_j,\beta)-\hat{h}(T_j,\beta))^{\otimes 2}+\sum\limits_{i=1}^{n}(g^{(1)}(X_i,\beta)-\tilde{h}(T_i,\beta))^{\otimes 2}\big]$, β^* satisfies $||\beta^*-\beta||\le||\hat{\beta}_{n,N}-\beta||$. We can easily prove that $C_{n,N}(\beta^*)\xrightarrow{p}\frac{\lambda}{1+\lambda}\Sigma_1(\beta)+\frac{1}{1+\lambda}\Sigma_2(\beta)$.

For $A(\beta)$, we have

$$\begin{aligned}\frac{1}{N}A(\beta) =&\frac{1}{N}\sum_{j=n+1}^{n+N}(Y_j - G(Z_j,\beta) - m(T_j))U(Z_j,\beta)\\ &+\frac{1}{N}\sum_{j=n+1}^{n+N}(G(Z_j,\beta) - \hat{G}(Z_j,\beta))U(Z_j,\beta)\\ &+\frac{1}{N}\sum_{j=n+1}^{n+N}(m(T_j) - \hat{m}(T_j,\beta))U(Z_j,\beta) + o_p(N^{-\frac{1}{2}})\\ :=&M_1 + M_2 + M_3 + o_p(N^{-\frac{1}{2}}).\end{aligned} \tag{17}$$

As the same argument of Liu [17], we can prove that

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}(G(Z_i,\beta) - g(X_i,\beta))U(Z_i,\beta) + o_p(n^{-\frac{1}{2}}). \tag{18}$$

Using the Kernel estimation method and Taylor expression, we have

$$\begin{aligned}M_3 =&\frac{1}{N}\sum_{j=n+1}^{n+N}(m(T_j) - \hat{m}(T_j,\beta))U(Z_j,\beta)\\ =&\frac{1}{N}\sum_{j=n+1}^{n+N}U(Z_j,\beta)m(T_j) - \frac{1}{N}\sum_{j=n+1}^{n+N}U(Z_j,\beta)\sum_{i=n+1}^{n+N}W_{Ni}(T_j)(m(T_i)+\varepsilon_i) + o_p(N^{-\frac{1}{2}})\\ =&\frac{1}{N}\sum_{j=n+1}^{n+N}U(Z_j,\beta)\sum_{i=n+1}^{n+N}W_{Ni}(T_j)(m(T_j) - m(T_i))\\ &-\frac{1}{N}\sum_{j=n+1}^{n+N}U(Z_j,\beta)\sum_{i=n+1}^{n+N}W_{Ni}(T_j)\varepsilon_i + o_p(N^{-\frac{1}{2}})\\ =&o_p(N^{-\frac{1}{2}}).\end{aligned} \tag{19}$$

This together with (17) and (18), we obtain that

$$\begin{aligned}\frac{1}{N}A(\beta) =&\frac{1}{N}\sum_{j=n+1}^{n+N}(Y_j - G(Z_j,\beta) - m(T_j))U(Z_j,\beta)\\ &+\frac{1}{n}\sum_{i=1}^{n}(G(Z_i,\beta) - g(X_i,\beta))U(Z_i,\beta) + o_p(N^{-\frac{1}{2}}).\end{aligned} \tag{20}$$

For $B(\beta)$, by simple calculation, it holds that

$$\begin{aligned}\frac{1}{n}B(\beta) =&\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i,\beta) - \tilde{m}(T_i,\beta))(g^{(1)}(X_i,\beta) - \tilde{h}(T_i,\beta))\\ =&\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i,\beta) - m(T_i))H(X_i,\beta) + o_p(n^{-\frac{1}{2}}).\end{aligned} \tag{21}$$

Then, we have

$$\begin{aligned} K(\beta) = & \frac{N}{n+N}\left[\frac{1}{N}\sum_{j=n+1}^{n+N}(Y_j - G(Z_j,\beta) - m(T_j))U(Z_j,\beta)\right. \\ & \left. + \frac{1}{n}\sum_{i=1}^{n}(G(Z_i,\beta) - g(X_i,\beta))U(Z_i,\beta)\right] \\ & + \frac{n}{n+N}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i,\beta) - m(T_i))H(X_i,\beta) + o_p(n^{-\frac{1}{2}})\right]. \end{aligned} \tag{22}$$

This together with (16), (20) and (21) complete the proof.

Remark 3 Obviously, compared to $\hat{\beta}_N$, $\hat{\beta}_{n,N}$ make full use of information, including response variable Y in the validation data, so it will give more accurate estimator than $\hat{\beta}_N$. This conclusion will be confirmed by simulation studies in the next section. However, in most applications, the primary data set is much larger than the validation data set, in such case, there is little information in the validation data, and this will lead to negligible difference between $\hat{\beta}_N$ and $\hat{\beta}_{n,N}$. On the other hand, $\hat{\beta}_N$ is simple for calculation. So, we recommend $\hat{\beta}_N$ when λ is large.

Clearly, the asymptotic covariances of $\hat{\beta}_N$ and $\hat{\beta}_{n,N}$ can be estimated by combining the sample moment method and the "plug-in" method. We give the following the notations:

$$\hat{\Sigma}_1(\beta) = \frac{1}{N}\sum_{j=n+1}^{n+N}(\hat{G}^{(1)}(Z_j,\beta) - \hat{h}^{(1)}(T_j,\beta))^{\otimes 2},$$

$$\hat{V}_0(\beta) = \frac{1}{N}\sum_{j=n+1}^{n+N}(Y_j - \hat{G}(Z_j,\beta) - \hat{m}(T_j,\beta))^2(\hat{G}^{(1)}(Z_j,\beta) - \hat{h}^{(1)}(T_j,\beta))^{\otimes 2},$$

$$\hat{V}_1(\beta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{G}(Z_i,\beta) - g(X_i,\beta))^2(\hat{G}^{(1)}(Z_i,\beta) - \hat{h}^{(1)}(T_i,\beta))^{\otimes 2},$$

$$\hat{\Sigma}_2(\beta) = \frac{1}{n}\sum_{i=1}^{n}(g^{(1)}(X_i,\beta) - \tilde{h}(T_i,\beta))^{\otimes 2},$$

$$\hat{V}_2(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i,\beta) - \tilde{m}(T_i,\beta))^2(g^{(1)}(X_i,\beta) - \tilde{h}(T_i,\beta))^{\otimes 2}.$$

Then, the asymptotic covariance of $\hat{\beta}_N$ and $\hat{\beta}_{n,N}$ can be consistently estimated by $\hat{\Sigma}_1^{-1}(\hat{\beta}_N)[\hat{V}_0(\hat{\beta}_N) + \lambda\hat{V}_1(\hat{\beta}_N)]\hat{\Sigma}_1^{-1}(\hat{\beta}_N)$ and $\hat{\Sigma}_3^{-1}(\hat{\beta}_{n,N})(\hat{V}(\hat{\beta}_{n,N}))\hat{\Sigma}_3^{-1}(\hat{\beta}_{n,N})$ with

$$\hat{V}(\hat{\beta}_{n,N}) = \frac{\lambda}{1+\lambda}(\hat{V}_0(\hat{\beta}_{n,N}) + \lambda\hat{V}_1(\hat{\beta}_{n,N})) + \frac{1}{1+\lambda}\hat{V}_2(\hat{\beta}_{n,N}),$$

respectively.

4 Simulation Results

In this section, we conducted some simulation studies to examine the finite sample performances of the proposed approaches.

To show the performance of the proposed estimators $\hat{\beta}_N$ and $\hat{\beta}_{N,n}$ in Section 2, we compared them with two other estimators: the naive estimator and the gold standard estimator. The naive estimator was obtained by ignoring the measurement error and applying the standard approach under model (1). The gold standard estimator consider all the true variable can be observed though it can not be obtained in practice.

The data are generated from the partially nonlinear model:

$$Y = g(X, \beta) + m(T) + e,$$

where $g(X, \beta) = 2\exp(-\beta X)$ with $\beta = 1$ and $m(T) = \sin(2\pi T)$ in which variables T is simulated from the uniform distribution on [0,1], X is measured with error and the surrogate variable $\tilde{X}$ is generated as $\tilde{X} = 1.25X + 0.2u$, X, e, u are standard normal distribution with truncation constants is 3, respectively. The simulation are run with validation data and primary data sizes of (n, N). The kernel function $K_1(x_1, x_2) = K_0(x_1)K_0(x_2)$ with $K_0(x) = (15/16)(1 - x^2)^2$ if $|x| \leq 1$, and 0 for otherwise. Let $K_2(x) = K_3(x) = K_0(x)$. Take the bandwidths $h_{1,n} = 0.2 * n^{-1/5}$, $h_{2,N} = 0.2 * N^{-1/5}$, $h_{3,n} = 0.2 * n^{-1/5}$, and truncation constant $b_n = 0.1 * n^{-1/42}$. To show the effects of the rate of the size of the primary data to the validation data, six cases are studied, which are $(n, N) = (60, 150), (120, 300), (30, 150), (60, 300), (30, 300), (60, 600)$, respectively. For each case, we replicated the simulation 1000 times. Table 1 presents the performance of four estimators of β. The 'mean' stands for the average of the 1000 estimates, and 'SD' is the standard deviation of the 1000 estimates.

Table 1: Means and deviations of $\hat{\beta}_N$, $\hat{\beta}_{n,N}$,$\hat{\beta}_{Naive}$ and $\hat{\beta}_{Gold}$ with different sample size

	Mean	SD	Mean	SD
$\lambda = 2.5$	$(n, N) = (60, 150)$		$(n, N) = (120, 300)$	
$\hat{\beta}_N$	0.8642	0.1998	1.0213	0.1552
$\hat{\beta}_{n,N}$	0.9583	0.0321	1.0004	0.0171
$\hat{\beta}_{Naive}$	0.7815	0.0213	0.7806	0.0158
$\hat{\beta}_{Gold}$	1.0002	0.0098	1.0000	0.0060
$\lambda = 5$	$(n, N) = (30, 150)$		$(n, N) = (60, 300)$	
$\hat{\beta}_N$	0.8595	0.2039	1.0272	0.1537
$\hat{\beta}_{n,N}$	0.9640	0.0740	1.0004	0.0344
$\hat{\beta}_{Naive}$	0.7817	0.0224	0.7810	0.0164
$\hat{\beta}_{Gold}$	0.9993	0.0101	1.0000	0.0066
$\lambda = 10$	$(n, N) = (30, 300)$		$(n, N) = (60, 600)$	
$\hat{\beta}_N$	0.8680	0.1529	1.0375	0.1192
$\hat{\beta}_{n,N}$	0.9513	0.0779	1.0052	0.0361
$\hat{\beta}_{Naive}$	0.7802	0.0178	0.7800	0.0128
$\hat{\beta}_{Gold}$	1.0004	0.0070	0.9999	0.0049

It follows from Table 1 that the naive estimators have much large bias than the gold standard estimators and the proposed estimators in all cases. The proposed estimators have a slight larger bias and SD than the gold standard estimators, which implies that the proposed estimators $\hat{\beta}_N$ and $\hat{\beta}_{n,N}$ work well. Compared with $\hat{\beta}_N$ and $\hat{\beta}_{n,N}$, $\hat{\beta}_{n,N}$ performs better than $\hat{\beta}_N$ in terms of that Mean is much close to the true value and SD is much smaller. This is caused by that the $\hat{\beta}_{n,N}$ involves more information in the estimation equation. But when the validation data sample is small, we suggest using $\hat{\beta}_N$, because it is much simple. The proposed estimation method performs well among different sample size of (n, N).

5 Conclusions

Nonlinear semiparametric model is a very useful semiparametric model which has been studied in many literatures. In this paper, we considered the situation of that the covariable is measured with error, furthermore, there is no specific structure assumption between the surrogate variable and the true variable. With the help of validation data, we obtain two estimators for unknown parameter in nonlinear function and prove its asymptotic normality, respectively. The first estimator is based on the primary data in (7) when applying the least squares method, moreover, the second estimator considers the response variable Y is available in the validation data as additional information in (11). The second estimator gives more accurate estimation at the cost of complexity. However, When the validation data sample is small and the primary data is large, there is little difference between these two estimators. In most cases, we recommend the first estimator because it is simple. Simulation studies show that the estimation methods we proposed are valid.

References

[1] Li R, Nie L. A new estimation procedure for partially nonlinear model via mixed-effects approach[J]. Canad. J. Stat., 2007, 35: 399–411.

[2] Li R, Nie L. Efficient statistical inference procedures for partially nonlinear models and their applications[J]. Biometrics, 2008, 64: 904–911.

[3] Huang T M, Chen H. Estimating the parametric component of nonlinear partial spline model[J]. J. Multivariate Anal., 2008, 99: 1665–1680.

[4] Song L X, Zhao Y, Wang X G. Sieve least squares estimation for partially nonlinear models[J]. Stat. Prob. Lett., 2010, 80: 1271–1283.

[5] Xiao Y T, Tian Z, Li F X. Empirical likelihood based inference for parameter and nonparametric function in partially nonlinear models[J]. J. Korean Stat. Soc., 2014, 43: 367–379.

[6] Li G R, Xue L G. Empirical likelihood confidence region for the parameter in a partially linear errors-in-variables model[J]. Comm. Stat. Theory Methods, 2008, 37: 1552–1564.

[7] Hu X M, Wang Z Z, Zhao Z W. Empirical likelihood for semiparametric varying-coefficient partially linear errors-in-variables models[J]. Stat. Prob. Lett., 2009, 79: 1044–1052.

[8] Wang X L, Li G R, Lin L. Empirical likelihood inference for semi-parametric varying-coefficient partially linear EV models[J]. Metrika., 2011, 73: 171–185.

[9] Wei C H, Mei C L. Empirical likelihood for partially linear varying-coefficient models with missing response variables and error-prone covariates[J]. J. Korean Stat. Soc., 2012, 41: 97–103.

[10] Wang Q H. Estimation of partial linear error-in-variables models with validation data[J]. J. Multivariate Anal., 1999, 69: 30–64.

[11] Wang Q H, Rao J N K. Empirical likelihood-based inference in linear errors-in-covariables models with validation data[J]. Biometrika, 2002, 89: 345–358.

[12] Stute W, Xue L G, Zhu L X. Empirical likelihood inference in nonlinear errors-in-covariables models with validation data[J]. J. Amer Stat. Assoc., 2007, 102: 332–346.

[13] Wang Q H, Zhang R Q. Statistical estimation in varying coefficient models with surrogate data and validation sampling[J]. J. Multivariate Anal., 2009, 100: 2389–2405.

[14] Du L L, Zou C L, Wang Z J. Nonparametric regression function estimation for error-in-variables models with validation data[J]. Stat. Sinica, 2011, 21: 1093–1113.

[15] Fang L D, Hu F X. Empirical likelihood dimension reduction inference in nonlinear EV models with validation data[J]. J. Math., 2012, 32: 113–120.

[16] Xue L G. Empirical likelihood inference in nonlinear semiparametric EV models with validation data[J]. Acta. Math. Sin., 2006, 49: 145–154.

[17] Liu Q. The estimation of nonlinear semiparametric EV models under missing data[J]. J. Sys. Sci. Math. Sci., 2010, 30: 1236–1250.

核实数据下非线性半参数EV模型的估计

肖燕婷[1,2], 田　铮[1], 孙　瑾[2]

(1.西北工业大学应用数学系, 陕西 西安 710129)

(2.西安理工大学应用数学系, 陕西 西安 710054)

摘要: 本文研究了核实数据下的协变量带有测量误差的非线性半参数EV模型. 在不假定测量误差结构的情形下, 利用最小二乘方法和核光滑技术, 构造了非线性函数中未知参数的两种估计, 证明了未知参数估计的渐近正态性. 通过数值模拟说明所提估计方法在有限样本下的有效性.

关键词: 非线性半参数EV 模型; 核实数据; 渐近正态性

MR(2010)**主题分类号:** 62G05　　**中图分类号:** O212.7