Open Journal of Statistics, 2015, 5, 808-819
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57080

Estimation of Nonparametric Multiple Regression Measurement Error Models with Validation Data

Zanhua Yin, Fang Liu

College of Mathematics and Computer Science, Gannan Normal University, Ganzhou, China
Email: yinzh226@163.com, yinzh226@nenu.edu.cn

Received 3 November 2015; accepted 27 December 2015; published 30 December 2015

Abstract

In this article, we develop estimation approaches for nonparametric multiple regression measurement error models when both independent validation data on covariables and primary data on the response variable and surrogate covariables are available. An estimator which integrates Fourier series estimation and truncated series approximation methods is derived without any error model structure assumption between the true covariables and surrogate variables. Most importantly, our proposed methodology can be readily extended to the case that only some of covariates are measured with errors with the assistance of validation data. Under mild conditions, we derive the convergence rates of the proposed estimators. The finite-sample properties of the estimators are investigated through simulation studies.

Keywords

Ill-Posed Inverse Problem, Linear Operator, Measurement Errors, Nonparametric Regression, Validation Data

1. Introduction

We can consider the following nonparametric regression model of a scaler response Y on an explanatory variable X

$$Y = g(X) + \varepsilon, \tag{1}$$

where $g(\cdot)$ is assumed to be a smooth, continuous but unknown nonparametric regression function and ε is a noise variable with $E(\varepsilon \mid X) = 0$ and $E(\varepsilon^2) < \infty$. It is not uncommon that the explanatory variable X is measured with error and instead only its surrogate variable W can be observed. In this case, one observes inde-

How to cite this paper: Yin, Z.H. and Liu, F. (2015) Estimation of Nonparametric Multiple Regression Measurement Error Models with Validation Data. *Open Journal of Statistics*, **5**, 808-819. http://dx.doi.org/10.4236/ojs.2015.57080

pendent replicates (W_i, Y_i), $1 \le i \le N$, of (W, Y) rather than (X, Y), where the relationship between W_i and X_i may or may not be specified. If not, the missing information for the statistical inference will be taken from a sample (W_j, X_j), $N+1 \le j \le N+n$, of so-call validation data independent of the primary (surrogate) sample. The objective of this manuscript is to estimate the unknown function $g(\cdot)$ via the surrogate data $\{(Y_i, W_i)\}_{i=1}^{N}$ and the validation data $\{(X_j, W_j)\}_{j=N+1}^{N+n}$.

A wide number of problems of similar type have attracted considerable attention in research literature over the past two decades (see, [1]-[6]). For instance, a quasi-likelihood method is intensively studied by [7]. A regression calibration approach is developed by [8] [9] and [10] [11] propose a method based on simulation-extrapolation (SIMEX) estimation. Other related methods include Bayesian approaches (see, [12]), semi-parametric method (see, [13] [14]), empirical likelihood method (see, [15]) and the instrumental variable method (see, [16]). Unfortunately, all these work mostly assume some parametric relationships between covariates and responses. Recently, nonparametric estimators of g have been developed by [17] and [18]. [17] develops a kernel-based approach for nonparametric regression function estimation with surrogate data and validation sampling. However, his method is not applicable for model (1) since it assumes that the response but not the covariable is measured with error. [18] proposes a nonparametric estimator which integrates local linear regression and Fourier transformation method when both explanatory and surrogate variable are scalars. Nonetheless, their method cannot be extended to multidimensional problems in which the explanatory variable vectors can consist of variables being measured with or/and without errors. For additional references and relevant topics for nonparametric regression models with measurement errors, ones may consult [19] and the references therein.

In practice, nonparametric estimation of g may not be an easy task since, as explained in Section 2, the relation that identifies g is a Fredholm equation of the first kind, *i.e.*

$$Tg = m, \tag{2}$$

which may lead to an ill-posed inverse problem. Ill-posed inverse problem related to nonparametric regression model has received considerable attention recently. [20] [21] consider kernel-based estimators while [22] and [23] develop series or sieve estimators. However, their methods require an instrumental variable, and assume that the explanatory variable X is directly observable without errors. In this article, we propose a nonparametric estimation approach which consists of two major steps. First, we propose estimators of generalized Fourier coefficients of T and m based on surrogate and validation data. Second, we replace the infinite-dimensional operator T by the finite-dimensional approximation to avoid higher-order coefficient estimation, and hence it develops an estimator of g. Furthermore, we extend this method to the case that only some of covariates are measured with errors. Under mild conditions, the consistencies of the resulting estimators are established and the convergence rates are also derived.

This article is arranged as follows. In Section 2, we first describe our estimation approach for the case that the covariates are all measured with errors. Extension to the case that only some of covariates are measured with errors will be discussed as well. We derive the convergence rates of our estimators under some regularity conditions in Section 3. Section 4 presents some numerical results from simulation studies. A brief discussion will be given in Section 5. Proofs of the theorems are presented in **Appendix**.

2. Methodology

We first describe our estimation approach for the case that the covariates are all measured with errors. In addition to the independent and identically distributed (i.i.d.) primary observations $\{(Y_i, W_i)\}_{i=1}^{N}$ from model (1), assume that i.i.d. validation data $\{(X_j, W_j)\}_{j=N+1}^{N+n}$ are also available. We shall suppose that X and W are both d-dimensional random vectors. Without loss of generality, let the supports of X and W both be contained in $\chi = [0,1]^d$ (otherwise, one can carry out monotone transformations of X and W).

In the following we let f_{XW}, f_X, f_W denote respectively the joint density of (X, W), marginal densities of X and W. Then we have

$$E(Y \mid W = w) = E[g(X) \mid W = w] = \int_{\chi} g(x) \frac{f_{XW}(x, w)}{f_W(w)} \mathrm{d}x. \tag{3}$$

According to Equation (3), g is actually the solution to an integral equation called Fredholm equation of the first kind. Let $m(w)=E(Y\mid W=w)f_W(w)$ and

$$L_2(\chi)=\left\{\varphi:\chi\to\mathcal{R},\ s.t.\ \|\varphi\|=\left(\int_\chi|\varphi(x)|^2\,\mathrm{d}x\right)^{1/2}<\infty\right\}.$$

Define the operator $T:L_2(\chi)\to L_2(\chi)$ by

$$(T\varphi)(w)=\int_\chi\varphi(x)\,f_{XW}(x,w)\,\mathrm{d}x.$$

Hence, Equation (3) is equivalent to the operator equation

$$m(w)=(Tg)(w). \tag{4}$$

For the unknown smooth function $g:\chi\to\mathcal{R}$, we assume that $g\in\mathcal{H}_s$ where

$$\mathcal{H}_s=\left\{g\in\mathcal{W}_2^s(\chi):\|g\|_{\mathcal{W}_2^s}<c\right\},$$

where c is a positive and finite constant. $\mathcal{W}_2^s(\chi)$ denotes the Sobolev space of smoothness $s\ge 1$, that is

$$\mathcal{W}_2^s(\chi)=\left\{\varphi\in L_2(\chi):\forall|\alpha|\le s,\partial_x^\alpha\varphi\in L_2(\chi)\right\},$$

where $\alpha=(\alpha_1,\cdots,\alpha_d)$, $|\alpha|=\alpha_1+\cdots+\alpha_d$, and the derivatives $\partial_x^\alpha\varphi=\dfrac{\partial^{|\alpha|}\varphi}{\partial x_1^{\alpha_1}\cdots x_d^{\alpha_d}}$. Given an integer s, the norm $\|g\|_{\mathcal{W}_2^s}$ is

$$\|g\|_{\mathcal{W}_2^s}=\left(\sum_{k=1}^{s}\left\|\partial_x^\alpha g\right\|^2\right)^{1/2},$$

here $\|\cdot\|$ denotes the norm on $L_2(\chi)$.

An estimator of g can then be obtained by replacing T and m by their series estimators based on surrogate data and validation data, and solving the resultant empirical version of (4). As before, let $\{\phi_k,k=1,2,\cdots\}$ denote a complete, orthonormal sequence for $L_2(\chi)$. Hence, we can write

$$m(w)=\sum_{k=1}^{\infty}m_k\phi_k(w),\ \text{and}\ \ f_{XW}(x,w)=\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}d_{kl}\phi_k(x)\phi_l(w),$$

where m_k and d_{kl} represent the generalized Fourier coefficients of m and f_{XW}, respectively. Intuitively, we can obtain the estimators of m_k, $m(w)$, d_{kl} and $f_{XW}(x,w)$ by

$$\hat{m}_k=\frac{1}{N}\sum_{i=1}^{N}Y_i\phi_k(W_i),\ \ \hat{m}(w)=\sum_{k=1}^{q}\hat{m}_k\phi_k(w),$$

$$\hat{d}_{kl}=\frac{1}{n}\sum_{j=N+1}^{N+n}\phi_k(X_j)\phi_l(W_j),\ \text{and}\ \ \hat{f}_{XW}(x,w)=\sum_{k=1}^{q}\sum_{l=1}^{q}\hat{d}_{kl}\phi_k(x)\phi_l(w),$$

respectively, where the integer q is a truncation point which is the main smoothing parameter in the approximating Fourier series. The operator T can then be consistently estimated by

$$(\hat{T}_n\varphi)(w)=\int_\chi\varphi(x)\,\hat{f}_{XW}(x,w)\,\mathrm{d}x.$$

Define the subset of $\mathcal{H}_s$:

$$\mathcal{H}_{ns}=\left\{\varphi=\sum_{k=1}^{q}\varphi_k\phi_k:\|\varphi\|_{\mathcal{W}_2^s}<c\right\}.$$

The estimator of $g(x)$ can be computed by

$$\hat{g}=\arg\min_{\varphi\in\mathcal{H}_{ns}}\left\|\hat{T}_n\varphi-\hat{m}\right\|^2. \tag{5}$$

Remark 1. *Let* $\tilde{W}_N$ *be the* $N\times q$ *matrix whose* (i,j) *element is* $\phi_j(W_i)$ *and* $\tilde{Y}_N=(Y_1,\cdots,Y_N)^{\mathrm{T}}$ *be the observed vector of Y based on the surrogate data* $\{(Y_i,W_i)\}_{i=1}^N$. *Let* $\bar{W}_n$ *and* $\bar{X}_n$, *respectively, denote the* $n\times q$ *matrices whose* (j,k) *elements are* $\phi_k(W_j)$ *and* $\phi_k(X_j)$ *based on the validation data. If* $A_n=\frac{1}{n}\bar{W}_n^{\mathrm{T}}\bar{X}_n$ *and* $b_N=\frac{1}{N}\tilde{W}_N^{\mathrm{T}}\tilde{Y}_N$, *then the solution to* (5) *assumes the following form*

$$\hat{g}(x)=\sum_{k=1}^{q}\hat{g}_k\phi_k(x) \tag{6}$$

where $\{\hat{g}_k,k=1,\cdots,q\}$ is given by $(\hat{g}_1,\cdots,\hat{g}_q)^{\mathrm{T}}=(A_n^{\mathrm{T}}A_n)^{-1}A_n^{\mathrm{T}}b_N$.

Next, we extend the estimator in (5) to nonparametric regression measurement error models with multi-covariates, that is

$$Y=g(X,Z)+\varepsilon, \tag{7}$$

where X is measured with error and W being its observed surrogate variable, and Z is measured without error. Let $\{(Y_i,W_i,Z_i)\}_{i=1}^N$ be a random sample from model (7), and $\{(X_j,W_j,Z_j)\}_{j=N+1}^{N+n}$ be i.i.d. validation observations. We assume that X and W are supported on $\chi=[0,1]^d$, and Z is supported on $[0,1]^p$.

Let $f_{XW|Z}$, $f_{X|Z}$ and $f_{W|Z}$ denote respectively the joint density of (X,W), marginal densities of X and W, all conditioning on $Z=z$. Similar to (3), for any $z\in[0,1]^p$, we have

$$m(w,z)=(T_z g)(w,z), \tag{8}$$

where $m(w,z)=E(Y\mid W=w,Z=z)f_{W|Z}(w)$, and the operator T_z is defined by

$$(T_z\varphi_z)(w,z)=\int_{\chi}\varphi_z(x)f_{XW|Z}(x,w)\mathrm{d}x,$$

where $\varphi_z=\varphi(\cdot,z)$ is any function on $L_2(\chi)$.

To obtain the estimator of $g(x,z)$, we set $K_h(u)=K(u/h)$ where K is a kernel function and $h>0$ is a bandwidth. Let $K_{p,h}(z)=\prod_{1\le k\le p}K_h(z_k)$. We consider the following estimators

$$\hat{m}_{zk}=\frac{(Nh_N^p)^{-1}\sum_{i=1}^{N}Y_i\phi_k(W_i)K_{p,h_N}(z-Z_i)}{(Nh_N^p)^{-1}\sum_{i=1}^{N}K_{p,h_N}(z-Z_i)},$$

and

$$\hat{d}_{zkl}=\frac{(nh_n^p)^{-1}\sum_{j=N+1}^{N+n}\phi_k(X_j)\phi_l(W_j)K_{p,h_n}(z-Z_j)}{(nh_n^p)^{-1}\sum_{j=N+1}^{N+n}K_{p,h_n}(z-Z_j)}.$$

Then we have

$$\hat{m}(w,z)=\sum_{k=1}^{q}\hat{m}_{zk}\phi_k(w),\ \text{ and }\ \hat{f}_{XW|Z}(x,w)=\sum_{k=1}^{q}\sum_{l=1}^{q}\hat{d}_{zkl}\phi_k(x)\phi_l(w).$$

Define the operator $\hat{T}_{nz}$ by

$$(\hat{T}_{nz}\varphi_z)(w,z)=\int_{\chi}\varphi_z(x)\hat{f}_{XW|Z}(x,w)\mathrm{d}x,$$

for any $\varphi_z\in L_2(\chi)$.

Then, for any $z\in[0,1]^p$, the estimator of $g(x,z)$ is

$$\tilde{g} = \arg\min_{\varphi_z \in \mathcal{H}_{ns}} \left\| \hat{T}_{nz}\varphi_z - \hat{m} \right\|^2. \tag{9}$$

Remark 2. *Denote* $f_N(z) = \left(Nh_N^p\right)^{-1} \sum_{i=1}^{N} K_{p,h_N}(z - Z_i)$ *and* $f_n(z) = \left(nh_n^p\right)^{-1} \sum_{j=N+1}^{N+n} K_{p,h_n}(z - Z_j)$. *Let* $\tilde{H}_N = f_N^{-1}(z) \times \operatorname{diag}\left(K_{p,h_N}(z - Z_1), \cdots, K_{p,h_N}(z - Z_N)\right)$ *and* $\bar{H}_n = f_n^{-1}(z) \times \operatorname{diag}\left(K_{p,h_n}(z - Z_{N+1}), \cdots, K_{p,h_n}(z - Z_{N+n})\right)$. *If* $\tilde{A}_n = \frac{1}{nh_n^p} \bar{W}_n^{\mathrm{T}} \bar{H}_n \bar{X}_n$ *and* $\tilde{b}_N = \frac{1}{Nh_N^p} \tilde{W}_N^{\mathrm{T}} \tilde{H}_N \tilde{Y}_N$, *then the solution to* (9) *has the following form*

$$\tilde{g}(x, z) = \sum_{k=1}^{q} \tilde{g}_{zk} \phi_k(x), \tag{10}$$

where $\{\tilde{g}_{zk}, k = 1, \cdots, q\}$ is given by $\left(\tilde{g}_{z1}, \cdots, \tilde{g}_{zq}\right)^{\mathrm{T}} = \left(\tilde{A}_n^{\mathrm{T}} \tilde{A}_n\right)^{-1} \tilde{A}_n^{\mathrm{T}} \tilde{b}_N$.

Remark 3. *If Z is discretely distributed with finite support, then* $g(x, z)$ *can be estimated by* (9) *with* $K_h(u)$ *being replaced by* $I(u = 0)$, *where* $I(\cdot)$ *is the indicator function.*

3. Theoretical Properties

In this section, we study the asymptotic properties of the estimators proposed in Section 2. We define τ_n (τ_{zn}) as a sieve measure of ill-posedness (see, [23]):

$$\tau_n = \sup_{\varphi \in \mathcal{H}_{ns}: \varphi \neq 0} \frac{\|\varphi\|}{\|T\varphi\|}; \quad \tau_{zn} = \sup_{\varphi_z \in \mathcal{H}_{ns}: \varphi_z \neq 0} \frac{\|\varphi_z\|}{\|T_z\varphi_z\|}.$$

First, we investigate the large-sample properties of the estimator $\hat{g}$. For this purpose, we present the following regular conditions which are mild and can be found in [24]) and [23].

A1. (i) The support of (X, W) is contained in χ^2; (ii) The joint probability measure of (Y, W) is absolutely continuous with respect to the product probability measure of *Y* and *W* and; (iii) The support of *W* is a cartesian product of compact connected intervals on which *W* has a probability density function that is bounded away from zero.

A2. For each $w \in \chi$, the function $E\left(Y^2 \mid W = w\right)$ is bounded by *c*.

A3. (i) $g \in \mathcal{H}_s$ with $g \in \mathcal{W}_2^s(\chi)$ and $s > 1$; and (ii) $m(W)$ belongs to $\mathcal{W}_2^r(\chi)$ with $r > 1/2$.

A4. The set of functions $\{\phi_k, k = 1, 2, \cdots\}$ is a orthonormal, complete basis for $L_2(\chi)$, and bounded uniformly over *k*.

A5. (i) $\lim n/N = \lambda$ for some constant $0 < \lambda \leq 1$; and (ii) $q = q(N, n) \to \infty$, $q/N \to 0$, $q/n \to 0$ as $n \to \infty$, $N \to \infty$.

Theorem 1. *Under conditions* A_1 - A_5, *as* $N \to \infty$ *and* $n \to \infty$, *we have*

$$\left\| \hat{g} - g \right\|_X = O_P\left\{ q^{-s/d} + \tau_n \times \left[q^{-r/d} + N^{-1/2} q^{1/2} + n^{-1/2} q^{1/2} \right] \right\}, \tag{11}$$

where $\|\cdot\|_X$ denotes $\|\varphi\|_X = \left\{ \int_\chi \varphi^2(x) f_X(x) \mathrm{d}x \right\}^{1/2}$ for any $\varphi \in L_2(\chi)$.

In (11), the term $q^{-s/d}$ arises from the bias of $\hat{g}$ caused by truncating the series approximation of *g*. The truncation bias decreases as *s* increases and *g* becomes smoother. Therefore, the smoother of *g* the faster the rate of convergence of $\hat{g}$. The terms $\tau_n N^{-1/2} q^{1/2}$ and $\tau_n n^{-1/2} q^{1/2}$ are respectively induced by random surrogate sampling errors and random validation sampling errors in the estimates of the generalized Fourier coefficients $\hat{g}_k$. When *X* is measured without error, the convergence rate of the sieve estimator of *g* is $O_P\left(q^{-s/d} + N^{-1/2} q^{1/2}\right)$. Comparing this rate to that in (11), we note that the bias part $q^{-s/d}$ is of the same order, however, the standard deviation part blows up from $N^{-1/2} q^{1/2}$ to $\tau_n \times \left[q^{-r/d} + N^{-1/2} q^{1/2} + n^{-1/2} q^{1/2} \right]$.

A more precise behaviour of the estimator can be obtained but depends on τ_n, as [23] discussed, which can be classified into mildly ill-posed case and severely ill-posed case. In the next corollary, we obtain these rates for the two particular cases.

Corollary 1. *Suppose the assumptions of Theorem* 1 *are satisfied.*

(i) Let $\tau_n = O\left(q^{(r-s)/d}\right)$ (mildly ill-posed case) with $r-s-1/2>0$, and $q \propto N^{d/(2r+d)}$, we have

$$\left\|\hat{g}-g\right\|_X = O_P\left(N^{-s/(2r+d)}\right);$$

(ii) Let $\tau_n = O\left\{q^{(r-s)/d}L(q)\right\}$ (severely ill-posed case) with $r-s-1/2>0$, and $q \propto N^{d/(2r+d)}$, we have

$$\left\|\hat{g}-g\right\|_X = O_P\left\{N^{-s/(2r+d)}L\left(N^{d/(2r+d)}\right)\right\},$$

where the function $L(q)$ goes to ∞ slowly such that $L(q)/q^{\varepsilon}\to 0$ for all $\varepsilon>0$.

Remark 4. *According to Corollary* 1(i), *the convergence rate becomes* $O\left(N^{-2/7}\right)$ *when* $r=3$, $s=2$ *and* $d=1$. *This is slower than that of the sieve estimator of a conditional mean function which can achieve the rate of convergence* $N^{-2/5}$.

Next, we study the large-sample properties of the estimator $\tilde{g}(x,z)$. For this purpose, we make the following assumptions.

B1. (i) The support of (X,W) is contained in χ^2, and Z is supported on $[0,1]^p$; (ii) Conditioning on $Z=z$, the joint probability measure of (Y,W) is absolutely continuous with respect to the product probability measure of Y and W and; (iii) Conditioning on $Z=z$, the support of W is a cartesian product of compact connected intervals on which W has a probability density function that is bounded away from zero.

B2. For each $(w,z)\in[0,1]^{d+p}$, $E\left(Y^2 \mid W=w, Z=z\right)$ is bounded by c.

B3. (i) For each $z\in[0,1]^p$, (8) has a solution $g(\cdot,z)\in\mathcal{H}_s$ with $g(\cdot,z)\in\mathcal{W}_2^s(\chi)$ and $s>1$ that does not depend on z and; (ii) For each $z\in[0,1]^p$, $m(W,Z=z)$ belongs to $\mathcal{W}_2^r(\chi)$ with $r>1/2$.

B4. (i) The set of functions $\{\phi_k, k=1,2,\cdots\}$ is a orthonormal, complete basis for $L_2(\chi)$, and bounded uniformly over k and; (ii) The kernel function K is a symmetrical, twice continuously differentiable function on $[-1,1]$, and $\int_{-1}^{1}u^j K(u)\mathrm{d}u=0$ for $j=1,\cdots,r-1$ and $\int_{-1}^{1}u^r K(u)\mathrm{d}u=c$, with $c\neq 0$ being some finite constant.

B5. (i) N, n, h_N, h_n satisfy the conditions that $Nh_N\to\infty$ and $nh_n\to\infty$; (ii) $h_n=c_h n^{-1/(2r+p)}$ and $h_N=C_h N^{-1/(2r+p)}$, where c_h and C_h are constants and $0<c_h,C_h<\infty$; and (iii) $q=c_q n^{\kappa d/(2r+d)}$ with $\kappa=2r/(2r+p)$ for some constant $c_q<\infty$.

B6. (i) $\lim n/N=\lambda$ for some constant $0<\lambda\le 1$; and (ii) $q=q(N,n)\to\infty$, $q/N\to 0$, $q/n\to 0$ as $n\to\infty$, $N\to\infty$.

Theorem 2. *Suppose assumptions* B_1 - B_6 *are satisfied. For each* $z\in[0,1]^p$, *let* $\tau_{zn}=O\left(q^{(r-s)/d}\right)$ *with* $r-s-1/2>0$, *we have*

$$\left\|\tilde{g}(x,z)-g(x,z)\right\|_{X|Z} = O_P\left[N^{-s\kappa/(2r+d)}\right].$$

The proofs of all the theorems are reported in **Appendix**.

4. Numerical Properties

In this subsection, we conducted a simulation study of the finite-sample performance of the proposed estimators. First, we choose the cosine sequence with $\phi_1(x)=1$ and $\phi_k(x)=\sqrt{2}\cos\left((k-1)\pi x\right), k=2,\cdots$ as the complete orthonormal basis for $L_2\left([0,1]\right)$, then get our estimators (denoted as $\hat{g}(x)$ and $\tilde{g}(x,z)$) following (6) and (10). For comparison, we consider [18] method (denoted as $\hat{g}_D$), and used the standard Nadaraya-Watson estimator with a Epanechnikov kernel to calculate $\hat{g}_N$ based on the primary dataset. It should be pointed out that $\hat{g}_N$ can serve as a gold standard in the simulation study, even though it is practically unachievable due to measurement errors. The performance of estimator g^{est} is assessed by using the average integrated squared errors (MISE) $\mathrm{MISE}=\frac{1}{M}\sum_{s=1}^{M}\left[g^{est}(u_s)-g(u_s)\right]^2$, where $u_s, s=1,\cdots,M$, are grid points at which $g^{est}(u_s)$ is evaluated.

Example 1: We considered model (1) with the regression function being

$$g(x)=\left[1-(2x-1)^2\right]^2 I_{\{x\in[0,1]\}},$$

and ε being distributed as $N(0,0.25)$. To perform this simulation, we generate X from a standard normal distribution, that is, $X\sim N(0,1)$, and assume that $W=\eta X+\left(1-\eta^2\right)^{1/2}v$, $v\sim N(0,1)$, and η is the standard deviation of the measurement error. Then, trim X and W in $[-2.5,2.5]$ and scale to $[0,1]$ respectively. Only results for $\eta=0.7$ and $\eta=0.9$ are reported here. Simulations were run with different validation and primary data sizes (n,N) ranging from $(10,30)$ to $(60,300)$ according to the ratio $\gamma=N/n=3$ and $\gamma=N/n=5$, respectively. For each case, 1000 simulated data sets were generated for each sample size of (n,N).

It is interesting to compare our estimator $\hat{g}$ with the estimators $\hat{g}_D$ and $\hat{g}_N$. Here, since our estimator $\hat{g}$ involves the regulation parameter q, we therefore present the following cross-validation (CV) selection criterion

$$\hat{q}_{CV}=\arg\min_q\left[\sum_{i=1}^{N}\left\{Y_i-\hat{g}^{(-i)}(W_i,q)\right\}^2\right],$$

where the subscript $-i$ meant that the estimator was constructed without using the ith observation (Y_i,W_i). For $\hat{g}_D$, [18] proposed an automatic way of choosing the smoothing parameters h_N, b_n and q. For $\hat{g}_N$, the CV approach is used for choosing bandwidth h_N.

Figure 1 shows the regression function curve $g(x)$, and the curves of the median MISEs based on 1000 replicated estimates of $\hat{g}$, $\hat{g}_D$ and $\hat{g}_N$ with $\eta=0.7$ under different sample size. From **Figure 1**, both $\hat{g}$ and $\hat{g}_D$ successfully capture the patterns of the true regression curves and have smaller bias than $\hat{g}_N$. As expected, $\hat{g}_N$ fails to produce accurate function curve estimates. In addition, it is obvious that the quality of our proposed estimator improve with the increase of sample sizes.

Table 1 compares, for various sample sizes, the results obtained for estimating curve $g(x)$ when $\eta=0.7$ or $\eta=0.9$. The estimated MISEs which were evaluated on a grid of 201 equidistant values of x in $[0,1]$ are presented. Our results show that the estimators $\hat{g}$ and $\hat{g}_D$ outperform $\hat{g}_N$. It is noteworthy that our proposed

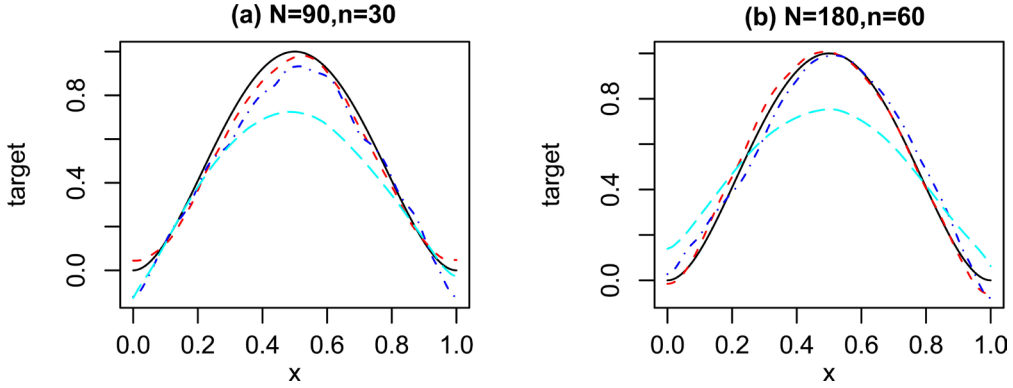

Figure 1. Curves for $\hat{g}(x)$, $\hat{g}_D(x)$ and $\hat{g}_N(x)$, and the regression function curve of $g(x)$. The solid, short-dashed, dash-dotted, and long-dashed curves respectively represent $g(x)$, $\hat{g}(x)$, $\hat{g}_D$ and $\hat{g}_N$.

Table 1. The estimated MISE ($\times 10^{-2}$) comparison for estimators $\hat{g}(x)$, $\hat{g}_D(x)$ and $\hat{g}_N(x)$ in **Example 1**.

	(n,N)	$\eta=0.7$			$\eta=0.9$		
		$\hat{g}(x)$	$\hat{g}_D(x)$	$\hat{g}_N(x)$	$\hat{g}(x)$	$\hat{g}_D(x)$	$\hat{g}_N(x)$
$\gamma=3$	(10, 30)	3.7858	5.1262	7.2254	3.8193	3.7822	6.8632
	(30, 90)	1.8630	2.2648	4.9030	1.3901	2.7834	5.0495
	(60, 180)	1.3056	1.8036	3.9296	0.8524	1.7082	3.7758
$\gamma=5$	(10, 50)	2.2956	3.1562	5.8640	2.0930	3.2097	5.2816
	(30, 150)	1.9711	2.1350	4.2684	1.6413	2.0299	4.0137
	(60, 300)	1.0628	1.5073	3.3890	0.7938	1.4438	3.2761

estimator generally performs better than the estimator proposed by [18] for the resultant MISEs of $\hat{g}$ are usually smaller. Also, the performance of $\hat{g}$ improves (*i.e.* the corresponding MISEs decrease) considerably as the sample sizes increases. For any nonparametric method in measurement error regression problem, the quality of the estimator also depends on the discrepancy of the observed sample. That is, the performance of the estimator depends on the variances of measurement error. Here, we compare the results for different values of η. As expected, **Table 1** shows that the effect of the variances on the estimator performance is obvious.

Example 2: We considered model (7) with the regression function being

$$g(x,z)=\left[1-(2x-1)^2\right]^2\exp(4x-1.4z-2)I_{\{(x,z)\in[0,1]^2\}},$$

and ε being distributed as $N(0,0.01)$. The covariate $(X,Z)^{\mathrm{T}}$ was generated from a bivariate normal distribution $N(0,\Sigma)$ with $var(X)=var(Z)=1$ and the correlation coefficient between *X* and *Z* being 0.6, and $W=\rho X+(1-\rho^2)^{1/2}v$, $v\sim N(0,1)$. Then, trim *X*, *W* and *Z* in $[-2.5,2.5]$ and scale to $[0,1]$ respectively. Results for $\rho=0.7$ and $\rho=0.9$ are reported. Simulations were run with different validation and primary data sizes (n,N) ranging from $(10,30)$ to $(60,300)$ according to the ratio $\gamma=N/n=3$ and $\gamma=N/n=5$, respectively. For each case, 1000 simulated data sets were generated for each sample size of (n,N).

Here, we only compared our estimator $\tilde{g}(x,z)$ with the naive estimator $\tilde{g}_N$ which is the multivariate kernel regression estimator based on the primary dataset $\{(Y_i,W_i,Z_i)\}_{i=1}^N$, since [18] method cannot be applied to multivariate cases. Here, we used the Epanechnikov kernel function $K(x)=0.75(1-x^2)$, $|x|\le 1$ for $\tilde{g}(x,z)$ and used an product kernel $K(x_1,x_2)=K_0(x_1)K_0(x_2)$ with $K_0(x)=-\frac{15}{8}x^2+\frac{9}{8},|x|\le 1$ for $\tilde{g}_N$. For the naive estimator $\tilde{g}_N$, bandwidth selection rules were considered by [25]. For our estimator $\tilde{g}(x,z)$, we used the cross-validation approach to choosing the three parameters h_N, h_n and *q*. For this purpose, h_n and (h_N,q) are selected separately as follows.

Define

$$\hat{f}_Z(z;h_n)=\frac{1}{nh_n}\sum_{j=N+1}^{N+n}K_{h_n}(z-Z_j).$$

Here, we adopt the cross-validation (CV) approach to estimate h_n by

$$\hat{h}_n=\arg\min_{h_n}\frac{1}{n}\sum_{j=N+1}^{N+n}\left\{Z_j-\hat{f}_Z^{(-j)}(Z_j;h_n)\right\}^2,$$

where the subscript $-j$ denotes the estimator being constructed without using the *j*th observation. After obtaining $\hat{h}_n$, we then select (h_N,q) by

$$(\hat{h}_N,\hat{q})=\arg\min_{h_N,q}\frac{1}{N}\sum_{i=1}^{N}\left\{Y_i-\tilde{g}^{(-i)}(W_i,Z_i;\hat{h}_n,h_N,q)\right\}^2,$$

where the subscript $-i$ denotes the estimator being constructed without using the *i*th observation (Y_i,W_i,Z_i).

We compute MISE at 101×101 grid points of (x,z) ranging in $[0,1]\times[0,1]$. **Table 2** reports the MISE for estimating curves $g(x,z)$ when $\rho=0.7$ or $\rho=0.9$ for various sample sizes. **Table 2** shows that our proposed estimator substantially outperformed the naive kernel estimator $\tilde{g}_N$. It is obvious that our proposed estimator $\tilde{g}$ has much smaller MISE than $\tilde{g}_N$.

5. Discussion

In this paper, we propose a new method for estimating non-parametric regression measurement error models using surrogate data and validation sampling. The covariates are measured with errors while we do not assume any error model structure between the true covariates and the surrogate variable. Most importantly, our proposed method can be readily extended to the multi-covariates model, say, $y=f(x,z)+\varepsilon$ where *x* is measured with error but *z* is measured exactly. Numerical results show that the new estimators are promising in terms of cor-

Table 2. The estimated MISE ($\times 10^{-2}$) comparison for the estimators $\tilde{g}(x,z)$ and $\tilde{g}_N(x,z)$ in **Example 2**.

	(n,N)	$\rho=0.7$		$\rho=0.9$	
		$\tilde{g}(x,z)$	$\tilde{g}_N(x,z)$	$\tilde{g}(x,z)$	$\tilde{g}_N(x,z)$
$\gamma=3$	(10, 30)	7.9858	9.8647	7.8340	9.4697
	(30, 90)	7.1677	9.0080	5.2114	8.0132
	(60, 180)	5.8270	8.7088	5.0759	6.7445
$\gamma=5$	(10, 50)	7.8902	9.7127	6.7415	9.7355
	(30, 150)	5.9597	9.2369	4.9424	7.0742
	(60, 300)	4.4316	8.3258	3.6832	6.9901

recting the bias arising from the errors-in-variables. It generally preforms better than the approach proposed by [18].

Acknowledgements

This work was supported by NSFC11301245, NSFC11501126 and Natural Science Foundation of Jiangxi Province of China under grant number 20142BAB211018.

References

[1] Pepe, M.S. and Fleming, T.R. (1991) A General Nonparametric Method for Dealing with Errors in Missing or Surrogate Covaraite Data. *Journal of the American Statistical Association*, **86**, 108-113. http://dx.doi.org/10.1080/01621459.1991.10475009

[2] Pepe, M.S. (1992) Inference Using Surrogate Outcome Data and a Validation Sample. *Biometrika*, **79**, 355-365. http://dx.doi.org/10.1093/biomet/79.2.355

[3] Lee, L.F. and Sepanski, J. (1995) Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data. *Journal of the American Statistical Association*, 90, 130-140. http://dx.doi.org/10.1080/01621459.1995.10476495

[4] Wang, Q. and Rao, J.N.K. (2002) Empirical Likelihood-Based Inference in Linear Errors-in-Covariables Models with Validation Data. *Biometrika*, **89**, 345-358. http://dx.doi.org/10.1093/biomet/89.2.345

[5] Zhang, Y. (2015) Estimation of Partially Linear Regression for Errors-in-Variables Models with Validation Data. *Springer International Publishing*, **322**, 733-742. http://dx.doi.org/10.1007/978-3-319-08991-1_76

[6] Xu, W. and Zhu, L. (2015) Nonparametric Check for Partial Linear Errors-in-Covariables Models with Validation Data. *Annals of the Institute of Statistical Mathematics*, **67**, 793-815. http://dx.doi.org/10.1007/s10463-014-0476-7

[7] Carroll, R.J. and Stefanski, L.A. (1990) Approximate Quasi-Likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, **85**, 652-663. http://dx.doi.org/10.1080/01621459.1990.10474925

[8] Carroll, R.J. and Wand, M.P. (1991) Semiparametric Estimation in Logistic Measurement Error Models. *Journal of the Royal Statistical Society*: *Series B*, **53**, 573-585.

[9] Sepanski, J.H. and Lee, L.F. (1995) Semiparametric Estimation of Nonlinear Errors-in-Variables Models with Validation Study. *Journal of Nonparametric Statistics*, **4**, 365-394. http://dx.doi.org/10.1080/10485259508832627

[10] Stute, W., Xue, L. and Zhu, L. (2007) Empirical Likelihood Inference in Nonlinear Errors-in-Covariables Models with Validation Data. *Journal of the American Statistical Association*, **102**, 332-346. http://dx.doi.org/10.1198/016214506000000816

[11] Cook, J.R. and Stefanski, L.A. (1994) Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, **89**, 1314-1328. http://dx.doi.org/10.1080/01621459.1994.10476871

[12] Carroll, R.J., Gail, M.H. and Lubin, J.H. (1993) Case-Control Studied with Errors in Covariables. *Journal of the American Statistical Association*, **88**, 185-199.

[13] Lü, Y.-Z., Zhang, R.-Q. and Huang, Z.-S. (2013) Estimation of Semi-Varying Coefficient Model with Surrogate Data and Validation Sampling. *Acta Mathematicae Applicatae Sinica, English Series*, **29**, 645-660. http://dx.doi.org/10.1007/s10255-013-0241-3

[14] Xiao, Y. and Tian, Z. (2014) Dimension Reduction Estimation in Nonlinear Semiparametric Error-in-Response Models with Validation Data. *Mathematica Applicata*, **27**, 730-737.

[15] Yu, S.H. and Wang, D.H. (2014) Empirical Likelihood for First-Order Autoregressive Error-in-Variable of Models with Validation Data. *Communications in Statistics Theory Methods*, **43**, 1800-1823. http://dx.doi.org/10.1080/03610926.2012.679763

[16] Stefanski, L.A. and Buzas, J.S. (1995) Instrumental Variable Estimation in Binary Regression Measurement Error Models. *Journal of the American Statistical Association*, **90**, 541-550. http://dx.doi.org/10.1080/01621459.1995.10476546

[17] Wang, Q. (2006) Nonparametric Regression Function Estimation with Surrogate Data and Validation sampling. *Journal of Multivariate Analysis*, **97**, 1142-1161. http://dx.doi.org/10.1016/j.jmva.2005.05.008

[18] Du, L., Zou, C. and Wang, Z. (2011) Nonparametric Regression Function Estimation for Error-in-Variable Models with Validation Data. *Statistica Sinica*, **21**, 1093-1113. http://dx.doi.org/10.5705/ss.2009.047

[19] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) Measurement Error in Nonlinear Models. Second Edition, Chapman and Hall CRC Press, Boca Raton. http://dx.doi.org/10.1201/9781420010138

[20] Hall, P. and Horowitz, J.L. (2005) Nonparametric Methods for Inference in the Presence of Instrumental Variables. *Annals of Statistics*, **33**, 2904-2929. http://dx.doi.org/10.1214/009053605000000714

[21] Darolles, S., Florens, J.P. and Renault, E. (2006) Nonparametric Instrumental Regression. Working Paper, GREMAQ, University of Social Science, Toulouse.

[22] Newey, W.K. and Powell, J.L. (2003) Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, **71**, 1565-1578. http://dx.doi.org/10.1111/1468-0262.00459

[23] Blundell, R., Chen, X. and Kristensen, D. (2007) Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica*, **75**, 1613-1669. http://dx.doi.org/10.1111/j.1468-0262.2007.00808.x

[24] Newey, W.K. (1997) Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics*, **79**, 147-168. http://dx.doi.org/10.1016/S0304-4076(97)00011-0

[25] Schimek, M.G. (2012) Variance Estimation and Bandwidth Selection for Kernel Regression. John Wiley & Sons, Inc., New York, 71-107.

[26] Timan, A. (1963) Theory of Approximation of Functions of a Real Variable. McMillan, New York.

Appendix

Proof of Theorem 1

Let $T^*: L_2(\chi) \to L_2(\chi)$ denotes the adjoint operator of T. Under assumption A1(ii), the self-adjoint operators of TT^* and T^*T have the same eigenvalue sequence $\{\mu_k^2\}$ with $\mu_1^2 = 1 \geq \mu_2^2 \geq \cdots$ Moreover, we assume that the corresponding eigenfunctions of the operators TT^* and T^*T are also orthonormal basis $\{\phi_k, k = 1, 2, \cdots\}$, and for all $k \geq 1$

$$T\phi_k = \mu_k \phi_k,\ T^*\phi_k = \mu_k \phi_k;\ \ T^*T\phi_k = \mu_k^2 \phi_k,\ TT^*\phi_k = \mu_k^2 \phi_k.$$

Define

$$g_n(x) = \sum_{k=1}^{q} g_k \phi_k(x), \text{ and } \quad m_N(w) = \sum_{k=1}^{q} m_k \phi_k(w).$$

Let T_n be the operator whose kernel is

$$t_n(x, w) = \sum_{k=1}^{q}\sum_{l=1}^{q} d_{kl}\phi_k(x)\phi_l(w),$$

then $T_n g_n = m_N$. By the definition of $\mathcal{H}_{ns}$, we have $g_n \in \mathcal{H}_{ns}$.

Lemma 1. *Under conditions* A1 *and* A3(i) *and the sieve space* $\mathcal{H}_{ns}$, *we have*

1) $\|T\{g - g_n\}\| \leq \text{const.} \times \mu_q \times \|g - g_n\|$;

2) $\tau_n \leq 1/\mu_q$.

Lemma 2. *Under conditions* A1, A3(ii) *and* A4, *we have*

$$\sup_{\varphi \in \mathcal{H}_{ns}} \left\| \left\{ \hat{T}_n - T \right\} \varphi \right\| = O_P\left(q^{-r/d} + q^{1/2} n^{-1/2} \right).$$

By some modifications of the proof of **Theorem 2** in [23] and applying the **Theorem 7** in [24], the proofs of **Lemma 1** and **Lemma 2** are straightforward and are omitted.

Proof of Theorem 1. By the triangle inequality, we have

$$\|\hat{g} - g\| \leq \|\hat{g} - g_n\| + \|g_n - g\|.$$

By the definition of $\mathcal{H}_{ns}$ and condition A3(i), we have

$$\|g_n - g\| = O\left(q^{-s/d}\right). \tag{12}$$

see e.g. [26] for Fourier series.

Next, by the definition of τ_n and the triangle inequality, we have

$$\|\hat{g} - g_n\| \leq \tau_n \left\| T(\hat{g} - g_n) \right\|.$$

We now analyze the term $\left\| T(\hat{g} - g_n) \right\|$. By the triangle inequality, we have

$$\begin{aligned} \left\| T(\hat{g} - g_n) \right\| &= \left\| \left(T - \hat{T}_n\right)\hat{g} + \hat{T}_n\hat{g} - \hat{m} + \hat{m} - m + T(g - g_n) \right\| \\ &\leq \left\| \left(T - \hat{T}_n\right)\hat{g} \right\| + \left\| \hat{T}_n\hat{g} - \hat{m} \right\| + \|\hat{m} - m\| + \left\| T(g_n - g) \right\|. \end{aligned}$$

By conditions A2, A4 and central limit theorem, we can show that $\|\hat{m} - E\hat{m}\| = O_P\left[(q/N)^{1/2} \right]$. From condition A3(ii), we have $\|E\hat{m} - m\| = O\left(q^{-r/d}\right)$. Hence, $\|\hat{m} - m\| = O_P\left(q^{-r/d} + q^{1/2} N^{-1/2}\right)$. In addition, by the definition of $\hat{g}$ and the triangle inequality, we have

$$\left\| \hat{T}_n\hat{g} - \hat{m} \right\| \leq \left\| \hat{T}_n g_n - \hat{m} \right\| \leq \left\| \left(\hat{T}_n - T\right) g_n \right\| + \left\| T(g_n - g) \right\| + \|m - \hat{m}\|.$$

These and **Lemma 2** imply

$$\|\hat{g} - g_n\| \leq \tau_n \times \left\{ O_P\left(q^{-r/d} + q^{1/2} N^{-1/2} + q^{1/2} n^{-1/2}\right) + O\left(\left\| T(g_n - g) \right\|\right) \right\}.$$

This and **Lemma 1** imply

$$\left\|\hat{g}-g\right\| \le \left\|g-g_n\right\| + \tau_n \times O_P\left(q^{-r/d} + q^{1/2}N^{-1/2} + q^{1/2}n^{-1/2}\right). \tag{13}$$

The theorem follows immediately from (12)-(13).

Proof of Theorem 2

Lemma 3. *For each* $z \in [0,1]^p$, *define*

$$g_{zn}(x) = \sum_{k=1}^{q} g_{zk}\phi_k(x), \quad \text{and} \quad m_{zN}(w) = \sum_{k=1}^{q} m_{zk}\phi_k(w).$$

Let T_{zn} be the operator whose kernel is

$$t_{zn}(x,w) = \sum_{k=1}^{q}\sum_{l=1}^{q} d_{zkl}\phi_k(x)\phi_l(w),$$

then $T_{zn}g_{zn} = m_{zN}$. By the definition of $\mathcal{H}_{ns}$, we have $g_{zn} \in \mathcal{H}_{ns}$.

Proof of Theorem 2. For each $z \in [0,1]^p$, by the triangle inequality, we have

$$\left\|\tilde{g}(x,z) - g(x,z)\right\| \le \left\|\tilde{g}-g_{zn}\right\| + \left\|g_{zn}-g\right\| \le \tau_{zn} \times \left\|T_z\left(\tilde{g}-g_{zn}\right)\right\| + \left\|g_{zn}-g\right\|,$$

By assumption B3(i), it is easy to show that $\left\|g_{zn}-g\right\| = O\left(q^{-s/d}\right)$.
Similar to the proof of Theorem 1, we have

$$\left\|T_z\left(\tilde{g}-g_{zn}\right)\right\| \le \left\|\left(T_z - \hat{T}_{zn}\right)\tilde{g}\right\| + \left\|\hat{T}_{zn}\tilde{g} - \hat{m}\right\| + \left\|\hat{m}-m\right\| + \left\|T_z\left(g-g_{zn}\right)\right\|.$$

According to assumptions B2, B3(ii), B4, and B5(i), we can show that $\left\|\hat{m}-E\hat{m}\right\| = O_P\left(q^{1/2}N^{-r/(2r+p)}\right)$, $\left\|E\hat{m}-m\right\| = O\left(q^{1/2}N^{-r/(2r+p)} + q^{-r/d}\right)$. In addition, by some modifications of the proof of Lemma 2, under assumptions B1, B3(ii), B4, B5(i) and B6, we have

$$\sup_{\varphi\in\mathcal{H}_{ns}} \left\|\left\{\hat{T}_{zn}-T\right\}\varphi\right\| = O_P\left(q^{-r/d} + q^{1/2}n^{-r/(2r+p)}\right).$$

For the term $\left\|T_z\left(g_{zn}-g\right)\right\|$, under assumptions B1, B3(i) and the sieve space $\mathcal{H}_{ns}$, we have

$$\tau_{zn} \times \left\|T_z\left(g_{zn}-g\right)\right\| \le \text{const.} \times \left\|g-g_{zn}\right\|$$

Combining all these results, we complete the proof. □