Ann. Inst. Statist. Math.
Vol. 55, No. 1, 21–39 (2003)


# ESTIMATION OF PARTIAL LINEAR ERROR-IN-RESPONSE MODELS WITH VALIDATION DATA*

QI-HUA WANG

*Institute of Applied Mathematics, Academy of Mathematics and System Science, Chinese Academy of Science, No. A1, Nansijie, Zhongguancun, Beijing 100080, China and School of Science, Heilongjiang University, Harbin 150000, China*

(Received May 23, 2000; revised September 3, 2001)

**Abstract.** In this paper, an estimation theory in partial linear model is developed when there is measurement error in the response and when validation data are available. A semiparametric method with the primary data is used to define two estimators for both the regression parameter and the nonparametric part using the least squares criterion with the help of validation data. The proposed estimators of the parameter are proved to be strongly consistent and asymptotically normal, and the estimators of the nonparametric part are also proved to be strongly consistent and weakly consistent with an optimal convergent rate. Then, the two estimators of the parameter are compared based on their empirical performances.

*Key words and phrases*: Partial linear model, validation data, strong consistency, asymptotic normality.

## 1. Introduction

In many research settings, the exact measurement of some important variables is difficult, time consuming, or expensive, and can only be performed for a few items in a large scale study. Hence surrogate data, which are more easily obtained using some relatively simple measuring methods, are used to measure these variables of interest. For example, in the evaluation of smoking behavior, current smoking behavior is generally collected by self-report using questionnaires. Data obtained by self-report are relatively inexpensive to obtain but may be subject to error. Expensive chemical analysis of saliva samples for the presence of contine can only be performed for at most a small subset of subjects enrolled in these large scale studies to yield a more accurate evaluation of smoking behavior. Analogous examples can be found in Wittes *et al.* (1989), Duncan and Hill (1985) and Pepe (1992) among others. Here, self-report of smoking behavior is used as a surrogate. The exact measurements obtained by expensive chemical analysis of saliva sample for a small subset of subjects together with their surrogate observations are usually treated as validation data set.

Generally, the relationship between the surrogate variables and the true variables can be rather complicated compared to the classical additive error model usually assumed (see, e.g., Fuller (1987)). In this case, some statisticians developed statistical inference techniques based on surrogate data and validation observations without specifying any

*Supported by NNSF of China (No. 10231030, No. 10241001) and a grant to the author for his excellent Ph.D. dissertation work in China.

error structure and the distribution assumption of the true variable given the surrogate variable (See, for example, Stefanski and Carroll (1987), Carroll and Wand (1991), Pepe and Fleming (1991), Pepe (1992), Pepe *et al.* (1994), Reilly and Pepe (1995), Sepanski and Lee (1995), and Wang (1999) and Wang and Rao (2002) among others). Carroll and Wand (1991) developed a semiparametric approach using the kernel regression technique for logistic measurement error models. Sepanski and Lee (1995) applied the method to nonlinear parametric model with error. Wang (1999) extended it to partial linear models with covariates measured erroneously. Also, Liang *et al.* (1999) considered the same problem under the additive error model.

The partial linear model is given by

$$Y = X^{\tau}\beta + g(T) + e, \tag{1.1}$$

where $Y$ is a scalar response variable, $X$ is a $p$-variate explanatory variable, $X^{\tau}$ is its transpose, $T$ is a random variable taking values in $[0,1]$, $\beta$ is a $p \times 1$ column vector of regression parameters, $g(\cdot)$ is an unknown regression function on $[0,1]$, $e$ is a random statistical error, and given $X$ and $T$ the errors $e = Y - X^{\tau}\beta - g(T)$ are assumed to be independent and identically distributed.

Such semiparametric additive models have been widely studied when $X$, $T$ and $Y$ are measured exactly. Various estimators for $\beta$ and $g(\cdot)$ have been proposed using various parametric and nonparametric methods such as the polynomial method (see, e.g., Heckman (1986), Rice (1986)), kernel method (see, e.g., Speckman (1988)), projection method (see, e.g., Chen (1988)) and nearest neighbor method (see, e.g., Hong (1991)). The estimation problem of $\beta$ and $g(\cdot)$ when $Y$ may be censored randomly on the right is considered by Wang (1996, 1997). It is well known that such censorship may occur in survival analysis, medical follow up and reliability studies.

In the present paper, we consider model (1.1) when $Y$ is measured with error and both explanatory variable $X$ and $T$ are measured exactly. That is, instead of the true $Y$, a surrogate variable $\tilde{Y}$ is observed. If the measurement error in the response is additive, the problem obviously reduces to the standard partially linear model and hence can be handled with standard methodology. Motivated by the above example and those in Duncan and Hill (1985) and Wittes *et al.* (1989), we consider the setting where some validation data are available to relate $Y$ and $\tilde{Y}$ and no error equation or distribution assumption of $Y$ given $\tilde{Y}$ is specified. In this case, estimators of both $\beta$ and $g(\cdot)$ are developed by using the kernel method and least squares method (see, e.g. Speckman (1988)) with the $Y_i$, that would have been used if measured exactly, replaced by a kernel regression estimate of $E[Y_i \mid \tilde{Y}_i, X_i, T_i]$ based on validation data. We define two estimators of $\beta$ and $g(\cdot)$ by two different ways of using data. For the definition of the first estimators, the least squares method uses only primary data except the use of validation data in the estimation of $E[Y_i \mid \tilde{Y}_i, X_i, T_i]$, where primary data include the surrogate data and the corresponding observations of $T$ and $X$. For the second estimators, the least squares method uses not only the primary data but also the true observations in the validation data set.

Asymptotic results for the two estimators are derived, showing that the two proposed estimators of $\beta$ are strongly consistent and asymptotically normal, and two estimators for $g(\cdot)$ are also strongly consistent and weakly consistent with an optimal convergent rate.

This paper is organized as follows. We define the two estimators of $\beta$ and $g(\cdot)$ in Section 2. The asymptotic results are formulated in Sections 3 and 4. In Section 5, we compare the two estimators of $\beta$ based on the theoretical framework as laid out in Sections 3 and 4. A simulation study is also carried out to compare them in Section 5. The proofs of the main results are presented in the Appendix.

## 2. Estimation

In what follows, assume that we have a primary data set containing independent and identically distributed observations of $\{(\tilde{Y}_j, X_j, T_j)_{j=n+1}^{n+N}\}$ and a validation data set containing $n$ indepedent and identically distributed observations of $\{(\tilde{Y}_i, Y_i, X_i, T_i)_{i=1}^{n}\}$. Also, it is assumed that the primary variables are independent of the validation variables and the random vector $(\tilde{Y}, X, T)$ in the primary data set and the one in the validation data set are identically distributed.

As in Carroll and Wand (1991), Pepe (1992) and Sepanski and Lee (1995) among others, measurements of the true response variables in the validation data set do not depend on $\tilde{Y}, X, T$. This is different from the case considered by Carroll *et al.* (1995) where true variables are observed with some probability $\pi$ which depends the observations of other relevant variables.

Since we do not make any assumptions on the relation between $Y$ and $\tilde{Y}$, the estimation problem of $\beta$ and $g(\cdot)$ in model (1.1) can not be handled directly with standard methodology. To use the surrogate data $\tilde{Y}$, it is necessary to rewrite the model (1.1) such that $\tilde{Y}$ is related to $X$ and $T$. For any matrix or vector $A$, denote by $A^\tau$ its transpose. Let $Z = (\tilde{Y}, X, T)$. Then the model (1.1) can be rewritten as

$$\begin{cases} E[Y \mid Z] = X^\tau \beta + g(T) + \epsilon \\ \epsilon = e - (Y - E[Y \mid Z]) \end{cases}. \tag{2.1}$$

Clearly, (2.1) is equivalent to model (1.1). Hence, statistical inference of $\beta$ and $g(\cdot)$ can be based on (2.1) instead of (1.1). Let $u(Z) = E[Y \mid Z]$. Indeed, the relation of $\tilde{Y}$, $X$ and $T$ is established in (2.1), but $u(z)$ is an unknown regression function. Hence, the model (2.1) can not be applied directly to get the estimators of $\beta$ and $g(\cdot)$ yet. If $u(Z) = \alpha^\tau Z$ is assumed, validation data set can be used to estimate parameter $\alpha$ and hence $u(\cdot)$ by least squares method. The results, however, can be sensitive to the specification of $u(z)$. Hence, here we do not assume any parametric structure. We use validation data to estimate $u(z)$ by the kernel regression estimation procedure. Hence, the semiparametric method due to Carroll and Wand (1991) and Sepanski and Lee (1995) can be extended to the model considered here to define the first estimators for $\beta$ and $g(\cdot)$.

By using the validation data, the regression function $u(z)$ in (2.1) can be estimated by a nonparametric kernel regression of $Y$ on $Z = z$. That is, the estimator of $u(Z)$ can be defined as

$$\hat{u}_n(z) = \frac{\sum_{i=1}^n Y_i K_1\left(\frac{Z_i - z}{b_n}\right)}{\sum_{i=1}^n K_1\left(\frac{Z_i - z}{b_n}\right)} \tag{2.2}$$

for any $z \in \tilde{\mathcal{Z}}$, where $\tilde{\mathcal{Z}}$ is the support set of $Z$, $K_1(\cdot)$ is a $p+2$ dimensional kernel function and $b_n$ a bandwidth tending to zero. It is well known that the choice of kernel-based method is common and natural to estimate regression. Other methods for regression estimation include the piece-wise constant smooth method, the method of smoothing spline, the orthogonal series approach and the local polynomial method. Local linear estimate, or more generally, local polynomial estimate, which was studied by Stone (1982), is known to have some favorable properties compared to the kernel method. Hence, kernel weighted local polynomial regression may be a better alternative in some sense. But, we will still use the kernel-based method here because it is simple and has desirable properties.

By (2.1), we have $u(Z) - X^\tau\beta = g(T) + \epsilon$. If $\beta$ were known, $g(\cdot)$ could then be estimated by the Nadaraya-Watson (N-W) kernel technique with $u(\cdot)$ replaced by $\hat{u}_n(\cdot)$. Let

$$W_{Nj}(t) = \frac{K_2\left(\frac{t-T_j}{h_N}\right)}{\sum_{j=n+1}^{n+N} K_2\left(\frac{t-T_j}{h_N}\right)},$$

where $K_2(\cdot)$ is also a kernel function and $h_N$ a bandwidth tending to zero. Using the above kernel method, we can define the first step estimator of $g(\cdot)$ as follows:

$$\begin{aligned}(2.3)\qquad \tilde{g}_{nN}(t) &:= \tilde{g}_{nN}(t,\beta) = \sum_{j=n+1}^{n+N} W_{Nj}(t)\hat{u}_n(Z_j) - \sum_{j=n+1}^{n+N} W_{Nj}(t)X_j^\tau\beta \\ &:= \hat{g}_{1,N}(t) - \hat{g}_{2,N}^\tau(t)\beta.\end{aligned}$$

Again by (2.1), we get $u(Z) - X^\tau\beta - g(T) = \epsilon$. Hence, the first estimator of $\beta$ is defined to be the one which minimizes $\hat{S}(\beta)$ given by

$$(2.4)\qquad \hat{S}(\beta) = \frac{1}{N}\sum_{j=n+1}^{n+N}(\hat{u}_n(Z_j) - X_j^\tau\beta - \hat{g}_{1,N}(T_j) + \hat{g}_{2,N}^\tau(T_j)\beta)^2.$$

That is, the estimator, say $\hat{\beta}$, minimizing (2.4) is the solution to the equation

$$(2.5)\qquad \frac{1}{N}\sum_{j=n+1}^{n+N}[(X_j - \hat{g}_{2,N}(T_j))(\hat{u}_n(Z_j) - \hat{g}_{1,N}(T_j) - (X_j - \hat{g}_{2,N}(T_j))^\tau\beta)] = 0.$$

By solving (2.5), it is easy to obtain that

$$(2.6)\qquad \hat{\beta} = \hat{\Sigma}^{-1}\hat{A},$$

where

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=n+1}^{n+N}(X_k - \hat{g}_{2,N}(T_k))(X_k - \hat{g}_{2,N}(T_k))^\tau,$$

$$\hat{A} = \frac{1}{N}\sum_{k=n+1}^{n+N}(X_k - \hat{g}_{2,N}(T_k))(\hat{u}_n(Z_k) - \hat{g}_{1,N}(T_k)).$$

By (2.3) and (2.6), we then obtain the first estimator of $g(\cdot)$ as follows

$$(2.7)\qquad \hat{g}(t) = \hat{g}_{1,N}(t) - \hat{g}_{2,N}^\tau(t)\hat{\beta}.$$

The above method first uses validation data to estimate $u(z)$. Then, only the primary data are employed to define the estimator of $\beta$ based on the least squares criterion after the one step estimator of $g(\cdot)$ is obtained. Such an approach uses the true observations on $Y$ only by the estimator of $u(z)$. The following procedure gives another alternative. Not only the validation data set is used in the estimation of $u(\cdot)$ but also the exact data in the validation data set together with primary data are used to define the estimator of $\beta$ with the least squares method.

Let

$$W_{ni}^*(t) = \frac{K^*\left(\frac{t-T_i}{h_n^*}\right)}{\sum_{i=1}^n K^*\left(\frac{t-T_i}{h_n^*}\right)},$$

$$\hat{g}_{1n}^*(t) = \sum_{i=1}^n W_{ni}^*(t) Y_i,$$

$$\hat{g}_{2n}^*(t) = \sum_{i=1}^n W_{ni}^*(t) X_i.$$

The second estimator of $\beta$ is defined to be the one which minimizes $\hat{S}^*(\beta)$ given by

$$\begin{aligned} \hat{S}^*(\beta) = &\frac{1}{N}\sum_{j=n+1}^{n+N}(\hat{u}_n(Z_j) - X_j^\tau\beta - \hat{g}_{1,N}(T_j) + \hat{g}_{2,N}^\tau(T_j)\beta)^2 \\ &+\frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\tau\beta - \hat{g}_{1,n}^*(T_i) + \hat{g}_{2,n}^{*^\tau}(T_i)\beta)^2. \end{aligned} \tag{2.8}$$

Similar to (2.6), it is obtained from (2.8) that the alternative estimator of $\beta$

$$\hat{\beta}^* = \hat{\Sigma}^{*^{-1}}\hat{A}^*, \tag{2.9}$$

where

$$\begin{aligned} \hat{\Sigma}^* = &\frac{1}{N}\sum_{k=n+1}^{n+N}(X_k - \hat{g}_{2,N}(T_k))(X_k - \hat{g}_{2,N}(T_k))^\tau \\ &+\frac{1}{n}\sum_{i=1}^n (X_i - \hat{g}_{2,n}^*(T_i))(X_i - \hat{g}_{2,n}^*(T_i))^\tau, \\ \hat{A}^* = &\frac{1}{N}\sum_{k=n+1}^{n+N}(X_k - \hat{g}_{2,N}(T_k))(\hat{u}_n(Z_k) - \hat{g}_{1,N}(T_k)) \\ &+\frac{1}{n}\sum_{i=1}^n (X_i - \hat{g}_{2,n}^*(T_i))(Y_i - \hat{g}_{1,n}^*(T_i)). \end{aligned}$$

(2.3) and (2.9) together then yield the second estimator of $g(\cdot)$ given by

$$\hat{g}^*(t) = \hat{g}_{1,N}(t) - \hat{g}_{2,N}^\tau(t)\hat{\beta}^*. \tag{2.10}$$

*Remark* 2.1. If one deletes the first term of the right hand side in (2.8) and obtains the estimator of $\beta$ by minimizing the second term of (2.8) only, the estimators of $\beta$ and $g(\cdot)$ are defined without using surrogate data. Intuitively, this will result in loss of some information. To make up for the loss of accuracy, one can instead increase $n$, the number of the observations of the exact data which are, however, expensive and difficult to obtain. Hence, this procedure is impractical though simple.

## 3. Asymptotic properties for $\hat{\beta}$ and $\hat{g}(t)$

The following Theorem 3.1 states the consistency of $\hat{\beta}$ and $\hat{g}(t)$.

THEOREM 3.1. *Suppose that all the assumptions listed in the Appendix* A *hold except* $[K^*]$ *and* $[h_n^*]$, *then*

$$\hat{\beta} \xrightarrow{a.s.} \beta. \tag{3.1}$$

*Furthermore, if* $\sup_{z\in\mathcal{Z}} E[Y^4 \mid Z = z] < \infty$, *then*

$$\hat{g}(t) \xrightarrow{a.s.} g(t) \tag{3.2}$$

*for any* $t \in [0,1]$.

Let $\mathcal{D}^m$ be the class of all continuous function $f$ on $R^{p+2}$ such that the derivatives

$$\frac{\partial^{i_1}}{\partial v_1^{i_1}}\frac{\partial^{i_2}}{\partial v_2^{i_2}}\cdots\frac{\partial^{i_{p+2}}}{\partial v_{p+2}^{i_{p+2}}}f(v_1,\dots,v_{p+2})$$

are uniformly bounded for $0 \le i_1 + i_2 + \cdots + i_{p+2} \le m$.

The following Theorem 3.2 states the asymptotic normality of $\hat{\beta}$ and gives the convergent rate of $\hat{g}(t)$.

THEOREM 3.2. *Under all the assumptions listed in the Appendix* B *except* $[h_n^*]'$, *if assumptions* $[X]$, $[u]$, $[Z]$ii, $[K_1]$ii, $[h_N]$ *and* $[Nn]$ *listed in the Appendix* A *are satisfied, we have*

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{V}), \tag{3.3}$$

*and*

$$\hat{g}(t) - g(t) = O_p((Nh_N^{3/2})^{-1}) + O_p((Nh_N)^{-1/2}) + O(h_N) + O_p(N^{-1/2}) \tag{3.4}$$

*for any* $t \in [0,1]$, *where*

$$\boldsymbol{V} = \Sigma^{-1}\boldsymbol{V}_1(\Sigma^{-1})^\tau,$$

$$\begin{aligned}\boldsymbol{V}_1 = {} & E[(u(Z) - X^\tau\beta - g(T))^2(X - E[X \mid T])(X - E[X \mid T])^\tau] \\ & + \lambda E[(Y - E[Y \mid Z])^2(X - E[X \mid T])(X - E[X \mid T])^\tau],\end{aligned}$$

$$\Sigma = E[(X - E[X \mid T])(X - E[X \mid T])^\tau]$$

*and* $\lambda$ *is defined in Assumption* $[Nn]'$.

*Remark* 3.1. The first component in the asymptotic covariance of $\hat{\beta}$ is the amount of information in the sample by modeling (1.1) as the regression relationship between $u(Z)$ and $X, T$. The second component is the extra dispersion caused by the nonparametric regression estimation of the unknown mean of $Y$ given $Z$ using the validation data set.

*Remark* 3.2. The asymptotic covariance of $\hat{\beta}$ can be estimated consistently by

$$\hat{V} = \widehat{\Sigma}^{-1}[\hat{\Lambda}_1 + \hat{\Lambda}_2](\widehat{\Sigma}^{-1})^\tau,$$

where $\hat{\Sigma}$ is that defined in Section 2 and

$$\hat{\Lambda}_1 = \frac{1}{N}\sum_{k=n+1}^{n+N}[(\hat{u}_n(Z_k) - X_k^\tau\hat{\beta} - \hat{g}_n(T_k))^2(X_k - \hat{g}_{2,N}(T_k))(X_k - \hat{g}_{2,N}(T_k))^\tau]$$

$$\hat{\Lambda}_2 = \frac{\lambda}{n}\sum_{i=1}^{n}[(Y_i - \hat{g}_{1n}(Z_i))^2(X_i - \hat{g}_{2,N}(T_i))(X_i - \hat{g}_{2,N}(T_i))^\tau].$$

COROLLARY 3.1. *Under all the assumptions of Theorem* 3.2, *if* $h_N = N^{-1/3}$ *we have*

$$\hat{g}(t) - g(t) = O_p(N^{-1/3}).$$

*Remark* 3.3. The convergent rate is the same as the optimal convergent rate for the corresponding nonparametric estimator of regression function (See, e.g. Stone (1980)).

## 4. Asymptotic properties for $\widehat{\beta}^*$ and $\hat{g}^*(t)$

We show in this section that $\widehat{\beta}^*$ and $\hat{g}^*(t)$ have the same asymptotic properties similar to $\hat{\beta}$ and $\hat{g}(t)$.

THEOREM 4.1. *Under the assumptions of Theorem* 3.1 *and* $[K^*]$ *and* $[h_n^*]$, *we have*

$$\hat{\beta}^* \xrightarrow{a.s.} \beta \tag{4.1}$$

*and*

$$\hat{g}^*(t) \xrightarrow{a.s.} g(t) \tag{4.2}$$

*for any* $t \in [0, 1]$.

THEOREM 4.2. *Under the assumptions of Theorem* 3.2 *and* $[K^*]$ *and* $[h_n^*]'$, *we have*

$$\sqrt{N}(\hat{\beta}^* - \beta) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{V}^*), \tag{4.3}$$

*and*

$$\hat{g}^*(t) - g(t) = O_p((Nh_N^{3/2})^{-1}) + O_p((Nh_N)^{-1/2}) + O(h_N) + O_p(N^{-1/2}) \tag{4.4}$$

*for any* $t \in [0,1]$, *where*

$$\begin{aligned}
\boldsymbol{V}^* &= (\Sigma^*)^{-1}\boldsymbol{V}_1^*(\Sigma^*)^{-\tau},\\
\boldsymbol{V}_1^* &= E[(u(Z) - X^\tau\beta - g(T))^2(X - E[X \mid T])(X - E[X \mid T])^\tau]\\
&\quad + \lambda\{E[(Y - E[Y \mid Z])^2(X - E[X \mid T])(X - E[X \mid T])^\tau]\\
&\qquad + E[(Y - X^\tau\beta - g(T))^2(X - E[X \mid T])(X - E[X \mid T])^\tau]\\
&\qquad + 2E[(Y - E[Y \mid Z])(Y - X^\tau\beta - g(T))(X - E[X \mid T])(X - E[X \mid T])^\tau]\},\\
\Sigma^* &= 2E[(X - E[X \mid T])(X - E[X \mid T])^\tau].
\end{aligned}$$

COROLLARY 4.1. *Under all the assumptions of Theorem* 4.2, *if* $h_N = N^{-1/3}$ *we have*

$$\hat{g}^*(t) - g(t) = O_p(N^{-1/3}).$$

*Remark* 4.1. It is intuitive that $\tilde{Y}$, $X$ and $T$ have useful information in predicting the unknown $Y$. Therefore, it is assumed that besides $Y$ and $\tilde{Y}$, $X$ and $T$ are also measured in the validation data set. Without observations on $X$, $T$ in the validation data set, one might define the estimators of $\beta$ and $g(\cdot)$ by rewriting the model (1.1) as

$$\begin{cases} u(\tilde{Y}) = X^\tau\beta + g(T) + \epsilon', \\ \epsilon' = e - (Y - u(\tilde{Y})), \\ u(\tilde{Y}) = E[Y \mid \tilde{Y}]. \end{cases} \tag{4.5}$$

If so, it is perhaps necessary to add the assumption $E[Y \mid Z] = E[Y \mid \tilde{Y}]$ to the corresponding theorems in Sections 3 and 4 in order to get the above asymptotic results for these estimators. Indeed, this can be seen by noticing that $E[Y \mid Z] = E[Y \mid \tilde{Y}]$ and $(C.e)i$ together imply that $E[\epsilon' \mid Z] = 0$, which is just needed in the proof of these results.

*Remark* 4.2. It is noted that the asymptotic covariances of $\hat{\beta}$ and $\hat{\beta}^*$ decrease if $\lambda$ decreases.

## 5. Simulation results

In this section, a simulation study was carried out to compare the two proposed estimators of $\beta$ with two naive estimator $\hat{\beta}_{Naive,1}$ and $\widehat{\beta}_{Naive,2}$ which are defined to be $\hat{\beta}$ and $\hat{\beta}^*$, respectively, with $\hat{u}_n(Z_k)$ replaced by $\tilde{Y}_k$ for $k = n+1, n+2, \ldots, n+N$.

We considered the partly linear model $Y = X\beta + g(T) + e$, where $\beta = 1.50$, $g(t) = 3.2t^2$ if $t \in [0,1]$, $g(t) = 0$ otherwise. It is assumed that $X$ and $e$ have a standard normal distribution and $T$ follows an uniform distribution on $[0,1]$. When $Y$ is subject to measurement error, validation data and primary data contain $(\tilde{Y}, X, T)$ and $(Y, \tilde{Y}, X, T)$, and $\tilde{Y}$ were generated by $\tilde{Y} = 1.12Y^2 + 0.85u$, where $u$ follows a standard normal distribution, and $u$ is independent of $e$. The simulation were run with validation data and primary data sizes of $(n, N) = (30,150), (60,300), (60,150), (120,300), (30,600)$ and $(60,1200)$, respectively. In the simulation study, $b_n$, $h_N$ and $h_n^*$ were taken to be $n^{-1/4}$, $N^{-1/2}$ and $n^{-1/2}$, and the kernel

$$K_1(x) = \begin{cases} -\dfrac{15}{8}x^2 + \dfrac{9}{8}, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Table 1. $\lambda = 5$.

| | $(n, N)$ | | | |
|---|---|---|---|---|
| | (30, 150) | | (60, 300) | |
| Estimate($\beta$) | Bias | SD | Bias | SD |
| $\hat{\beta}_{Naive,1}$ | −1.0275 | 2.4148 | −0.9823 | 2.0316 |
| $\hat{\beta}_{Naive,2}$ | −1.1651 | 1.9571 | −1.1262 | 1.5852 |
| $\hat{\beta}$ | 0.1097 | 1.4021 | 0.0984 | 1.0721 |
| $\hat{\beta}^*$ | 0.1109 | 1.0428 | 0.0995 | 0.7305 |

Table 2. $\lambda = 2.5$.

| | $(n, N)$ | | | |
|---|---|---|---|---|
| | (60, 150) | | (120, 300) | |
| Estimate($\beta$) | Bias | SD | Bias | SD |
| $\hat{\beta}_{Naive,1}$ | −1.1210 | 2.1027 | −1.0011 | 1.6542 |
| $\hat{\beta}_{Naive,2}$ | −1.1081 | 1.7034 | −1.1732 | 1.2731 |
| $\hat{\beta}$ | 0.0995 | 1.1527 | 0.0471 | 0.7648 |
| $\hat{\beta}^*$ | 0.0977 | 0.8023 | 0.0615 | 0.5262 |

Table 3. $\lambda = 20$.

| | $(n, N)$ | | | |
|---|---|---|---|---|
| | (30, 600) | | (60, 1200) | |
| Estimate($\beta$) | Bias | SD | Bias | SD |
| $\widehat{\beta}_{Naive,1}$ | −1.2117 | 1.7820 | −1.2841 | 0.1363 |
| $\widehat{\beta}_{Naive,2}$ | −1.3024 | 1.5827 | −1.3245 | 0.8577 |
| $\hat{\beta}$ | 0.1073 | 0.8758 | 0.0577 | 0.5129 |
| $\hat{\beta}^*$ | 0.1088 | 0.8243 | 0.0595 | 0.4317 |

of order $m = 4$ and the kernel

$$K_2(x) = \begin{cases} \frac{15}{16}(1 - 2x^2 + x^4), & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

were used to calculate $u_n(\cdot)$ and the weight $W_{Nj}(t)$, $j = n+1, \ldots, n+N$, respectively. For calculating $W^*_{ni}(t)$, $1 = 1, 2, \ldots, n$, $K^*(x)$ were taken to be same as $K_2(x)$. The simulation results are presented in following Tables 1, 2 and 3 to compare the bias and standard deviation (SD) of $\hat{\beta}$ and $\hat{\beta}^*$ with $\widehat{\beta}_{Naive,1}$ and $\widehat{\beta}_{Naive,2}$. The bias and variance are computed from 1000 simulated data sets of each size $(n, N)$. That is, 1000 estimates were computed to yield the results in the tables.

From the simulation results, the two naive estimators have serious bias and bigger SD than the two proposed estimators. Comparing the two proposed estimators, they have approximately the same bias but different standard deviation. $\hat{\beta}^*$ behaves better than $\hat{\beta}$ in the sense of less standard deviation. But, it seems that the performances of the estimators are close to each other when $n$, $N$ and $\lambda$ are large enough. Hence, it is also appropriate to use $\hat{\beta}$ for practical purposes because it is simpler than $\hat{\beta}^*$ and it also has little bias.

On the other hand, the simulation study illustrates that the change of size of validation set yields bigger effect on the proposed estimators than that of the surrogate set in terms of bias and SD.

## 6. Concluding remarks

It is noted that the dimension of $Z = (\tilde{Y}, X, T)$ is $p+2$. When $p$ is large, the curse of dimensionality may occur because of the kernel estimation of $u(z) = E[Y \mid Z = z]$. For this case, one solution is to consider this dimension reduction model by assuming $u(z) = m(\alpha^\tau z)$, where $\alpha$ is a $(p+2) \times 1$ vector of unknown parameter and $m(\cdot)$ is an unknown function. Actually, this assumed model is a single index model, and hence can be expected to apply in a wide variety of situation. To estimate $u(z)$, one can first estimate $\alpha$ by sliced inverse regression techniques due to Li (1991), and then estimate $m(\cdot)$ by the kernel regression technique with univariate explanatory variable. After obtaining the estimator of $u(z)$, we can develop the estimate theory for $\beta$ and $g(\cdot)$ and obtain the results similar to those obtained in previous sections. The asymptotic inference theory in this setting is worth further investigating.

## Appendix A: Proofs of Theorems 3.1 and 4.1

To prove Theorems 3.1 and 4.1, we first introduce the following notations, assumptions and present several lemmas.

Let $g_1(t) = E[u(Z) \mid T = t]$, $g_2(t) = E[X \mid T = t]$. Denote by $g_{2r}(\cdot)$ and $X_{ir}$ the $r$-th component of $g_2(\cdot)$ and $X_i$, $i = 1, 2, \ldots, n$; $r = 1, 2, \ldots, p$. Let $\|a - b\| = \sum |a_i - b_i|$ for any vectors $a$ and $b$, where $a_i$ and $b_i$ are the $i$-th component of $a$ and $b$, respectively.

The following assumptions are needed for the strong consistency of $\hat{\beta}$ and $\hat{g}(t)$.

$[g]$: $g_1(t)$, $g_{2r}(t)$ and $g(t)$ satisfy Lipschitz condition of order 1, $r = 1, 2, \ldots, p$.

$[r]$: The density of $T$, say $r(t)$, exists and satisfies

$$0 < \inf_{0 \le t \le 1} r(t) \le \sup_{0 \le t \le 1} r(t) < \infty.$$

$[X]$: $\sup_t E[X_{1r}^4 \mid T = t] < \infty, r = 1, 2, \ldots, p$.

$[u]$: There exists absolute constant $c_0$ such that

$$|u(z_1) - u(z_2)| \le c_0 \|z_1 - z_2\|.$$

$[Z]$i: The density of $Z$, say $f_Z(z)$, exists and satisfies

$$\sum_{N=1}^{\infty} NP(f_Z(Z) < \eta_N) < \infty$$

for some positive constant sequence $\eta_N > 0$ tending to zero.

ii: $f_Z(z)$ has bounded partial derivative of order one.

$[e]$i: $E[e \mid X, T] = 0$,

ii: $\sup_{z \in \mathcal{Z}} E[e^2 \mid Z = z] < \infty$,

$[Y]$: $\sup_{z \in \mathcal{Z}} E[Y^2 \mid Z = z] < \infty$ for some $\delta > 0$.

$[K_1]$i: $K_1(\cdot)$ is a bounded kernel function with bounded support.
ii: $K_1(\cdot)$ is a kernel of order one .
iii: $|K_1(\cdot)|$ is of bounded variation.
$[K_2]$: There exists absolute constants $M_1, M_2$ and $\rho > 0$ such that

$$M_1 I[|t| \le \rho] \le K_2(t) \le M_2 I[|t| \le \rho].$$

$[h_N]$i: $\sum_{N=1}^{\infty} N^{-1} h_N < \infty$.
ii: For $\eta_N$ appearing in $[Z]$i, $\sum_{N=1}^{\infty}(\eta_N^2 N h_N^{5/2})^{-2} < \infty$.
$[Nn]$: $\frac{N}{n} = O(1)$.
$[h_N b_n]$: $\frac{b_n^{p+2}}{h_N} = O(1)$.

The above assumptions are needed for the proof of Theorem 3.1. To prove Theorem 4.1, we futher assume the following conditions.
$[K^*]$: There exists absolute constants $M_1^*$, $M_2^*$ and $\rho^* > 0$ such that

$$M_1^* I[|t| \le \rho^*] \le K^*(t) \le M_2^* I[|t| \le \rho^*].$$

$[h_n^*]$: $\sum_{n=1}^{\infty} n^{-1} h_n^* < \infty$ and $\sum_{n=1}^{\infty}(n^2 h_n^*)^{-1} < \infty$.

*Remark* A.1. Assumption $[Z]$i is clearly satisfied when $\inf_{z \in \mathcal{Z}} f_Z(z) = \eta$ and $\eta_N$ is taken to be $\frac{1}{2}\eta$ for some $\eta > 0$. Assumption $[Z]$i also holds if $(\tilde{Y}, X)$ follows $p+1$-dimensional standard normal distribution, $T$ follows uniform distribution on $[0,1]$ and $\eta_N$ is taken to be $\frac{1}{N^{2/(p+1)} \log^2 N}$. Analogously, we also can give some examples for the assumption $[Z]'$i listed in the Appendix B.

LEMMA A.1. *Under assumptions* $[r]$ *and* $[K_2]$, *we have as* $Nh_N \to \infty$
(a) $E[W_{Nj}(T_i)]^\gamma \le c(N^\gamma h_N)^{-1}$, $\gamma = 2, 4$; $i, j = n+1, \ldots, n+N$.
(b) $E[W_{Nj}(t)]^\gamma \le c(N^\gamma h_N)^{-1}$, $\gamma = 2, 4$; $j = n+1, \ldots, n+N$.

The Lemma is proved by Qin (1995) and Wang (1996).

LEMMA A.2. *Under conditions* $[g]$, $[K_2]$, $[r]$ *and* $[X]$, *we have*

$$\hat{\Sigma} \xrightarrow{a.s.} \Sigma.$$

The proof is similar to that of Hong (1991).
For the sake of simplicity, let us denote by $c$ any positive constant.

PROOF OF THEOREM 3.1. Clearly

$$\hat{\beta} - \beta = \hat{\Sigma}^{-1} \tilde{A}(\beta) \tag{A.1}$$

where $\hat{\Sigma}$ is as defined in Section 2, and

$$\tilde{A}(\beta) = \frac{1}{N} \sum_{k=n+1}^{n+N} (X_k - \hat{g}_{2,N}(T_k))[\hat{u}_n(Z_k) - \hat{g}_{1,N}(T_k) - (X_k - \hat{g}_{2,N}(T_k))^\tau \beta].$$

To prove Theorem 3.1 by Lemma A.2, it is sufficient to prove that

$$\tilde{A}(\beta) \xrightarrow{a.s.} 0. \tag{A.2}$$

For simplicity, we next denote $\sum_k$ and $\sum_{k,j}$ to be the summations on $k$ and on $k$ and $j$, respectively, extending over the integers from $n+1$ to $n+N$. Analogious definition also applies to $\sum_j$. From (2.1), $\tilde{A}(\beta)$ can be decomposed as:

$$\begin{aligned}\tilde{A}(\beta) &= \frac{1}{N}\sum_k (X_k - \hat{g}_{2,N}(T_k))(\hat{u}_n(Z_k) - u(Z_k)) \\ &+\frac{1}{N}\sum_k (X_k - \hat{g}_{2,N}(T_k))\epsilon_k \\ &+\frac{1}{N}\sum_k (X_k - \hat{g}_{2,N}(T_k))(g(T_k) - \sum_j W_{Nj}(T_k)(u(Z_j) - X_j^\tau \beta)) \\ &-\frac{1}{N}\sum_k (X_k - \hat{g}_{2,N}(T_k))\sum_j W_{Nj}(T_k)(\hat{u}_n(Z_j) - u(Z_j)) \\ &:= A_{N,1} + A_{N,2} + A_{N,3} + A_{N,4}.\end{aligned} \tag{A.3}$$

To prove (A.2), it is sufficient to prove $A_{N,i} \xrightarrow{a.s.} 0, i = 1,2,3,4$, by (A.3). We prove only the case: $i = 1$. The proofs of the other three cases are much easier and hence are omitted.

Next, We consider $A_{N,1}$. Recalling the definition of $\hat{g}_{2,N}(\cdot)$, it follows that

$$\begin{aligned}A_{N,1} &= \frac{1}{N}\sum_{k,j} W_{Nj}(T_k)(E[X_j \mid T_j] - X_j)(\hat{u}_n(Z_k) - u(Z_k)) \\ &+\frac{1}{N}\sum_{k,j} W_{Nj}(T_k)(g_2(T_k) - g_2(T_j))(\hat{u}_n(Z_k) - u(Z_k)) \\ &+\frac{1}{N}\sum_k (X_k - E[X_k \mid T_k])(\hat{u}_n(Z_k) - u(Z_k)) \\ &:= A_{N,11} + A_{N,12} + A_{N,13}.\end{aligned} \tag{A.4}$$

By the definition of $\hat{u}_n(\cdot)$, $A_{N,11}$ can be represented as

$$\begin{aligned}A_{N,11} &= \frac{1}{nNb_n^{p+2}}\sum_{k,j} W_{Nj}(T_k)(E[X_j \mid T_j] - X_j) \\ &\times\frac{\sum_{i=1}^n (Y_i - E[Y_i \mid Z_i])K_1\left(\frac{Z_i - Z_k}{b_n}\right)}{\hat{f}_Z(Z_k)} I\left[\hat{f}_n(Z_k) \geq \frac{1}{2}f_Z(Z_k) \geq \frac{1}{2}\eta_N\right] \\ &+\frac{1}{nNb_n^{p+2}}\sum_{k,j} W_{Nj}(T_k)(E[X_j \mid T_j] - X_j) \\ &\times\frac{\sum_{i=1}^n (u(Z_i) - u(Z_k))K_1\left(\frac{Z_i - Z_k}{b_n}\right)}{\hat{f}_Z(Z_k)} I\left[\hat{f}_n(Z_k) \geq \frac{1}{2}f_Z(Z_k) \geq \frac{1}{2}\eta_N\right]\end{aligned} \tag{A.5}$$

$$
\begin{aligned}
&+\frac{1}{N}\sum_{k,j} W_{Nj}(T_k)(E[X_j \mid T_j] - X_j)(\hat{u}_n(Z_k) - u(Z_k)) \\
&\times I\left[\hat{f}_n(Z_k) < \frac{1}{2} f_Z(Z_k), f_Z(Z_k) \geq \eta_N\right] \\
&+\frac{1}{N}\sum_{k,j} W_{Nj}(T_k)(E[X_j \mid T_j] - X_j)(\hat{u}_n(Z_k) - u(Z_k)) \\
&\times I[f_Z(Z_k) < \eta_N] \\
&\triangleq A_{N,11}^{(1)} + A_{N,11}^{(2)} + A_{N,11}^{(3)} + A_{N,11}^{(4)}.
\end{aligned}
$$

Denote by $A_{N,11}^{[i,r]}$ the $r$-th component of $A_{N,11}^{(i)}$, $i = 1,2,3,4$. By conditions $[Y]$, $[Z]$i, $[K_1]$i, $[g]$, $[K_2]$, $[X]$ and $[h_N b_n]$ together with Lemma A.1, we have

$$
\begin{aligned}
\text{(A.6)}\quad E[A_{N,11}^{[1,r]}]^2 &\leq \frac{n}{N(nb_n^{p+2})^2\eta_N^2}\sum_{k,j} E\{W_{Nj}^2(T_k)E[(X_{jr} - E[X_{jr} \mid T_j])^2 \mid T_j]\} \\
&\leq c(nb_n^{(3/2)(p+2)}\eta_N)^{-2}.
\end{aligned}
$$

Using conditions $[h_N]$ii, $[Nn]$ and $[h_N b_n]$, we get

$$
\text{(A.7)}\qquad A_{N,11}^{[1,r]} \xrightarrow{a.s.} 0, \qquad r = 1,2,\ldots,p.
$$

From $[u]$, $[K_1]$i and $[Z]$i, we have

$$
\begin{aligned}
|A_{N,11}^{[2,r]}| &\leq \frac{b_n}{\eta_N N}\sum_k \left[\left|\sum_j W_{Nj}(T_k)(E[X_{jr} \mid T_j] - X_{jr})\right| \right. \\
&\qquad \left. \times \frac{\frac{1}{nb_n^{p+2}}\sum_{i=1}^n \left\|\frac{Z_i - Z_k}{b_n}\right\| \left|K_1\left(\frac{Z_i - Z_k}{b_n}\right)\right|}{\hat{f}_n(Z_k)}\right] \\
&\leq \frac{c}{\eta_N N b_n^{p+1}}\sum_k \left|\sum_j W_{Nj}(T_k)(E[X_{j_r} \mid T_{j_r} - X_{jr})\right|, \qquad r = 1,2,\ldots,p.
\end{aligned}
$$

By $[X]$, $[Z]$i, $[K_1]$i, $[h_N b_n]$ and Lemma A.1, similar to (A.6) it follows

$$
\text{(A.8)}\qquad E[A_{N,11}^{[2,r]}]^4 \leq c(N^2 h_N^5 \eta_N^4)^{-1}, \qquad r = 1,2,\ldots,p.
$$

By condition $[h_N]$ii, we get

$$
\text{(A.9)}\qquad A_{N,11}^{[2,r]} \xrightarrow{a.s.} 0, \qquad r = 1,2,\ldots,p.
$$

For any $\epsilon > 0$, we have

$$
P(|A_{N,11}^{[3,r]}| > \epsilon) \leq \sum_k P\left(|\hat{f}_Z(Z_k) - f_Z(Z_k)| > \frac{1}{2}\eta_N\right) \leq cb_n^{-2p+1}\exp\{-cn\eta_N b_n^{p+2}\}
$$

by $[Z]$i. By condition $[Nn]$ and $\sum_{n=1}^{\infty} nb_n^{-2p+1}\exp\{-cn\eta_N b_n^{p+2}\} < \infty$ (this is implied by $[h_N]$ii, $[Nn]$ and $[h_N b_n]$, and Borel-Cantelli's Lemma), it follows that

$$A_{N,11}^{[3,r]} \xrightarrow{a.s.} 0, \qquad r = 1, 2, \ldots, p. \tag{A.10}$$

Clearly, for any $\epsilon > 0$ we have

$$P(|A_{N,11}^{[4,r]}| > \epsilon) \le P\left(\bigcup_{k=n+1}^{n+N} \{f_Z(Z_k) < \eta_N\}\right) \le NP(f_Z(Z) < \eta_N).$$

This together with $[Z]$i proves $\sum_{N=1}^{\infty} P(|A_{N,11}^{[4,r]}| > \epsilon) < \infty$. By Borel-Cantelli's Lemma, we have

$$A_{N,11}^{[4,r]} \xrightarrow{a.s.} 0, \qquad r = 1, 2, \ldots, p. \tag{A.11}$$

Combining (A.5), (A.7), (A.9), (A.10) and (A.11), it follows that $A_{N,11}$ and similarly $A_{N,1i}, i = 2, 3$, converge to zero almost surely. This proves $A_{N,1} \xrightarrow{a.s.} 0$.

The proofs of (A.2) and hence (3.1) are completed by Lemma A.2.

Next, we prove (3.2). Clearly

$$\hat{g}(t) - g(t) = \hat{g}_{1,N}(t) - g_1(t) - (\hat{g}_{2,N}^{\tau}(t) - g_2^{\tau}(t))\hat{\beta} + g_2^{\tau}(t)(\hat{\beta} - \beta). \tag{A.12}$$

Standard arguments can be used to prove $\hat{g}_{1,N}(t) \xrightarrow{a.s.} g_1(t)$ and $\hat{g}_{2,N}(t) \xrightarrow{a.s.} g_2(t)$. This together with (3.1) proves (3.2).

The proof of Theorem 3.1 is thus completed. Similarly, we can prove Theorem 4.1.

## Appendix B: Proofs of Theorems 3.2 and 4.2

To prove the asymptotic normality of $\hat{\beta}$ and obtain the rate of weak consistency, the assumption $[X]$ can be weaken to

$[X]'$: $\sup_t E[X_{1r}^2 \mid T = t] < \infty, r = 1, 2, \ldots, p$.

And the assumptions $[u]$, $[Z]$, $[K_1]$ii, $[h_N]$ and $[Nn]$ can be replaced by

$[u]'$: For some $m > p + 2$, $u(\cdot) \in \mathcal{D}^m$.

$[Z]'$i: The density of $Z$, say $f_Z(z)$, exists and satisfies $NP(f_Z(Z) < \eta_N) \to 0$ for some positive constant sequence $\eta_N > 0$ tending to zero.

ii: $f_Z(z) \in \mathcal{D}^m$.

$[K_1]'$: $K_1$ is a kernel of order $m$.

$[h_N]'$ i: $Nh_N^2 \to \infty$,

ii: $Nh_N^4 \to 0$.

$[Nn]'$: $\frac{N}{n} \to \lambda$, where $\lambda$ is a nonnegative constant.

$[\Sigma]'$: $E[(X - E[X \mid T])(X - E[X \mid T])^{\tau}$ is a positive definite matrix.

$[h_N b_n]'$i: For $\eta_N$ appearing in $(C.Z)'$i, $\eta_N^2 nb_n^{2(p+2)} \to \infty$,

ii: $\frac{nb_n^{2m}}{h_N\eta_N^2} \to 0$ for $m$ appearing in $[u]'$.

The above assumptions are for the proof of Theorem 3.2. For Theorem 4.2, we need further assume the following conditions in addition to the above assumptions.

$[h_n^*]'$i: $nh_n^* \to \infty$.

ii: $nh_n^{*4} \to 0$.

Let $U_k = (\epsilon_k, Z_k), V_i = (Y_i, Z_i), k = 1, 2, \ldots, n + N; i = 1, 2, \ldots, n$.

LEMMA B.1. *Under the conditions of Theorem 3.2, we have*

$$\sqrt{N}\tilde{A}(\beta) = \frac{1}{n\sqrt{N}} \sum_{k=n+1}^{n+N} \sum_{i=1}^{n} \Psi_{n1}(U_k, V_i; b_n) + o_p(1),$$

*where*

$$\Psi_{n1}(U_k, V_i; b_n) = (X_k - E[X_k \mid T_k])\epsilon_k + \frac{(X_k - E[X_k \mid T_k])(Y_i - u(Z_k))K_1\left(\frac{Z_i - Z_k}{b_n}\right)}{b_n^{p+2} f(Z_k)},$$

$i = 1, 2, \ldots, n, k = n+1, \ldots, n+N$.

PROOF. Let $\hat{l}(z) = \frac{1}{nb_n^{p+2}} \sum_{i=1}^{n} Y_i K_1\left(\frac{Z_i - z}{b_n}\right)$. Then

$$\begin{aligned}
\sqrt{N}\tilde{A}(\beta) &= \frac{1}{\sqrt{N}} \sum_k (X_k - E[X_k \mid T_k]) \frac{\hat{l}(Z_k) - \hat{f}(Z_k)u(Z_k)}{f(Z_k)} \\
&+ \frac{1}{\sqrt{N}} \sum_k (X_k - E[X_k \mid T_k])\epsilon_k \\
&+ \sqrt{N}A_{N,11} + \sqrt{N}A_{N,12} + \sqrt{N}A_{N,131} + \sqrt{N}A_{N,22} \\
&+ \sqrt{N}A_{N,23} + \sqrt{N}A_{N,3} + \sqrt{N}A_{N,4}
\end{aligned} \tag{B.1}$$

where $A_{N,11}, A_{N,12}$ are defined in (A.4), $A_{N,3}$ and $A_{N,4}$ are defined in (A.3) and

$$A_{N,22} = \frac{1}{N} \sum_{k,j} W_{Nj}(T_k)(g_2(T_k) - g_2(T_j))\epsilon_k$$

$$A_{N,23} = -\frac{1}{N} \sum_{k,j} W_{Nj}(T_k)(X_j - E[X_j \mid T_j])\epsilon_k$$

$$A_{N,131} = \frac{1}{\sqrt{N}} \sum_k (X_k - E[X_k \mid T_k]) \frac{(\hat{l}(Z_k) - \hat{f}_Z(Z_k)u(Z_k))(f_Z(Z_k) - \hat{f}_Z(Z_k))}{f_Z(Z_k)\hat{f}_Z(Z_k)}.$$

Recalling the definitions of $\hat{l}(\cdot)$ and $\hat{f}(\cdot)$, the sum of the first two terms in (B.1) is clearly the main term of the asymptotic representation in Lemma B.1. Hence, to prove Lemma B.1, it is sufficient to prove that the last seven terms on the right hand side of (B.1) converge to zero in probability. Using the analogous arguments to treat $A_{N,11}$ in the Appendix A, we can prove $\sqrt{N}A_{N,131} \xrightarrow{p} 0$. Next, we prove $\sqrt{N}A_{N,11} \xrightarrow{p} 0$ only. The other five terms can be proved to converge to zero similarly.

Next, let us consider $A_{N,11}$.

By (A.6) and conditions $[h_N b_n]$, $[h_N b_n]'$i, it follows that

$$\sqrt{N}A_{N,11}^{[1,r]} \xrightarrow{p} 0, \qquad r = 1, 2, \ldots, p. \tag{B.2}$$

Let

$$\zeta_n(Z_k) = \frac{1}{nb_n^{p+2}} \sum_{i=1}^{n} (u(Z_i) - u(Z_k))K_1\left(\frac{Z_i - Z_k}{b_n}\right).$$

By $[Z]$i, we have

$$(\text{B.3}) \qquad E(A_{N,11}^{[2,r]})^2 \leq \frac{1}{N\eta_N^2}\sum_k E\left\{\left(\sum_j W_{Nj}(T_k)(E[X_j \mid T_j] - X_j)\right)^2\right\}$$
$$\times[E(\zeta_n(Z_k) - E[\zeta_n(Z_k) \mid Z_k])^2 + E(E^2[\zeta_n(Z_k) \mid Z_k])],$$

where $A_{N,11}^{[2,r]}$ is defined in (A.5). Notice that

$$(\text{B.4}) \quad E[\zeta_n(Z_q) \mid Z_q] = \int (u(Z_q + b_n u) - u(Z_q))f(Z_q + b_n u)K_1(u)du$$
$$= \int P_n(u)K(u)du + h_n^k \int R_n(\xi, u)K_1(u)du, \qquad 1 < \xi < 1$$

where $P_n(u)$ is a polynomial of degree $k-1$ on $u$ and hence $\int P_n(u)K_1(u)du = 0$ and $R_n(\xi, n)$ is the $k$-th remainder of the Taylor's expansion and satisfies $\int R_n(\xi, u)K(u)du < \infty$ by $[K_1]'$. This proves

$$(\text{B.5}) \qquad |E[\zeta_n(Z_q) \mid Z_q]| \leq cb_n^m, \qquad q = n+1, \ldots, n+N,$$

where $c$ is a constant which does not dependent on $Z_q$.

Clearly, by derivative mean-value Theorem, we have

$$(\text{B.6}) \qquad E[(\zeta_n(Z_k) - E[\zeta_n(Z_k) \mid Z_k])^2]$$
$$\leq \frac{1}{nb_n^{2(p+2)}} \int_{\mathcal{Z}} (u(Z) - u(Z_k))^2 K_1^2\left(\frac{z - Z_k}{b_n}\right) f_Z(z)dz$$
$$\leq \frac{c}{nb_n^{2p}} \to 0.$$

By (B.3)–(B.6) and Lemma A.1 together with $[h_N b_n]'$, we get

$$(\text{B.7}) \qquad NE(A_{N,11}^{[2,r]})^2 \leq c\left(\frac{1}{nb_n^{2(p+2)}\eta_N^2} + \frac{b_n^{2m}}{h_N\eta_N^2}\right) \to 0$$

which proves that $\sqrt{N}A_{N,11}^{[2,r]} \xrightarrow{p} 0, r = 1, 2, \ldots, p$. Similar to (A.10) and (A.11), we can prove that $\sqrt{N}A_{N,11}^{[i,r]} \xrightarrow{p} 0, r = 1, 2, \ldots, p$; $i = 3, 4$. Hence, This together with (A.5) and (B.2) proves that $\sqrt{N}A_{N.11} \xrightarrow{p} 0$.

The proof of Lemma B.1 is thus completed.

LEMMA B.2. *Under the conditions of Theorem* 3.2, *we have*

$$\frac{1}{n\sqrt{N}} \sum_{i=1}^{n} \sum_{k=n+1}^{n+N} \Psi_{n1}(U_k, V_i; b_n) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{V}_1)$$

*where* $\boldsymbol{V}_1$ *is defined in Theorem* 3.2.

PROOF. For any $p$-dimension vector $\alpha$, let

$$\phi_n(U,V;b_n) = \alpha^\tau \Psi_{n1}(U,V;b_n)$$
$$M_n = \frac{1}{n\sqrt{N}} \sum_{i=1}^{n} \sum_{k=n+1}^{n+N} \phi_n(U_k,V_i;b_n).$$

Clearly, $M_n$ is a two sample statistic and

$$E[\phi_n(U,V;b_n) \mid V] \to \alpha^\tau (X - E[X \mid T])(Y - E[Y \mid Z]). \tag{B.8}$$

Similarly, by $[u]'$ and $[K_1]'$, we have

$$\begin{aligned} E[\phi_n(U,V;b_n) \mid U] &= \alpha^\tau (X - E[X \mid T])\epsilon \\ &\quad + \frac{\alpha^\tau (X - E[X \mid T]) \int (u(z) - u(Z)) K_1 \left(\frac{z-Z}{b_n}\right) dz}{b_n^{p+2} f_Z(z)} \\ &\to \alpha^\tau (X - E[X \mid T])\epsilon. \end{aligned} \tag{B.9}$$

By the first equality in (B.9), we get

$$\begin{aligned} E[\phi_n(U,V;b_n)] &= E\{E[\phi_n(U,V;b_n) \mid U]\} \\ &= E\frac{\alpha^\tau (X - E[X \mid T]) \int (u(z) - u(Z)) K_1 \left(\frac{z-Z}{b_n}\right) dz}{b_n^{p+2} f_Z(z)}. \end{aligned} \tag{B.10}$$

Similar to (B.5), it follows that

$$|E[\phi_n(U,V;b_n)]| \le c b_n^m. \tag{B.11}$$

Hence, by $[Nn]'$ and $[h_N b_n]'ii$, we have

$$\sqrt{N} E\phi_n(U,V;b_n) \to 0. \tag{B.12}$$

Lemma B.1 of Sepanski and Lee (1995) together with (B.9), (B.10) and (B.12) proves that

$$\frac{1}{n\sqrt{N}} \sum_{i=1}^{n} \sum_{k=n+1}^{n+N} \phi_n(U_k,V_i;b_n) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{V}_2)$$

where

$$\begin{aligned} \boldsymbol{V}_2 &= E[(u(z) - X\beta - g(T))^2 \alpha^\tau (X - E[X \mid T])(X - E[X \mid T])^\tau \alpha] \\ &\quad + \lambda E[(Y - E[Y \mid Z])^2 \alpha^\tau (X - E[X \mid T])(X - E[X \mid T])^\tau \alpha]. \end{aligned}$$

This proves Lemma B.2.

PROOF OF THEOREM 3.2. Notice that

$$\sqrt{N}(\hat{\beta} - \beta) = \sqrt{N}\Sigma^{-1}\tilde{A}(\beta) + \sqrt{N}(\hat{\Sigma}^{-1} - \Sigma^{-1})\tilde{A}(\beta). \tag{B.13}$$

By Lemma B.1 and B.2, it follows that

$$\sqrt{N}\tilde{A}(\beta) = O_p(1). \tag{B.14}$$

Hence, (B.13) and (B.14) together with Lemmas A.2 and B.1, B.2 prove (3.3).

In what follows, we prove (3.4).

Using the analogous arguments as before, we can prove that $G_{N1} = O_p((Nh_N^{3/2})^{-1}) + O_p((Nh_N)^{-1/2})$, $G_{Ni} = O_p((Nh_N)^{-1/2})$, $i = 2, 3$ and $EG_{N4} = O(h_N)$ where $G_{N1}$, $G_{N2}$, $G_{N3}$ and $G_{N4}$ are as defined in (A.15). This together with (A.15) yields that

$$\hat{g}_{1,N}(t) - g_1(t) = O_p((Nh_N^{3/2})^{-1}) + O_p((Nh_N)^{-1/2}) + O(h_N). \tag{B.15}$$

Similarly, we can prove that

$$\hat{g}_{2,N}(t) - g_2(t) = O_p((Nh_N)^{-1/2}) + O(h_N). \tag{B.16}$$

The result (3.3) implies that

$$\hat{\beta} - \beta = O_p(N^{-1/2}). \tag{B.17}$$

Formulas (A.14), (B.15)-(B.17) together prove (3.4).

This completes the proof of Theorem 3.2. Similarly, we can prove Theorem 4.2.

## References


Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measure error models, *J. Roy. Statist. Soc., Ser. B*, **53**, 652–663.

Carroll, R. J., Knickerbocker, R. K. and Wang, C. Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates, *Ann. Statist.*, **23**, 161–181.

Chen, H (1988). Convergence rates for parametric components in a partly linear model, *Ann. Statist.*, **16**, 136–146.

Duncan, G. and Hill, D. (1985). An investigations of the extent and consequences of measurement error in labor-economics survey data, *Journal of Labor Economics*, **3**, 508–532.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, New York.

Heckman, N. (1986). Spline smoothing in partly linear models, *J. Roy. Statist. Soc. Ser. B*, **48**, 244–248.

Hong, S. Y. (1991). Asymptotic properties of kernel smoothing in partial linear model, *The Fourth China-Japan Symposium on Statist. Kuming*, 131–134.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion), *J. Amer. Statist. Assoc.*, **86**, 337–342.

Liang, H., Härdle, W. and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model, *Ann. Statist.*, **27**, 1519–1535.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample, *Biometrika*, **79**, 355–365.

Pepe, M. S. and Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data, *J. Amer. Statist. Assoc.*, **86**, 108–113.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method, *J. Statist. Plann. Inference*, **42**, 137–160.

Qin, Q. S. (1995). Kernel smoothing method in partial linear model with random censored data, *Chinese Ann. Math. Ser. A*, **16**(4), 441–453.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models, *Biometrika*, **82**, 299–314.

Rice, J. (1986). Convergence rates for partially splined models, *Statist. Probab. Lett.*, **4**, 203–208.

Sepanski, J. H. and Lee, L. F. (1995). Semiparametric estimation of nonlinear error-in-variables models with validation study, *J. Nonparametr. Statist.*, **4**, 365–394.

Speckman, P. (1988). Kernel smoothing in partial linear models, *J. Roy. Statist. Soc. Ser. B*, **50**, 413–436.

Stefanski, L. A. and Carrol, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement error models, *Biometrika*, **74**, 703–716.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators, *Ann. Statist.*, **8**, 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regresion, *Ann. Statist.*, **10**, 1040–1053.

Wang, Q. H. (1996). Consistent estimates in random censorship semiparametric models, *Sci. China Ser. A*, **39**, 163–176.

Wang, Q. H. (1997). Some asymptotic properties of the estimators in semiparametric regression models with randomly censored data, *Sci. China Ser. A*, **40**, 945–957.

Wang, Q. H. (1999). Estimation of partial linear error-in-variables model, *J. Multivariate Anal.*, **69**, 30–64.

Wang, Q. H. and Rao, J. N. K. (2002). Empirical likelihood-based inference in linear errors-in-covariables models with validation data, *Biometrica*, **89**, 345–358.

Wittes, J., Lakatos, E. and Probstfied, J. (1989). Surrogate endpoints in clinical trails: Cardiovascular diseases, *Statistics in Medicine*, **8**, 415–425.