INVESTIGACIONES ECONOMICAS. Volumen XIX (3) - Septiembre 1995, pp. 435-467

ON ASYMPTOTIC INFERENCES IN NONPARAMETRIC AND SEMIPARAMETRIC MODELS WITH DISCRETE AND MIXED REGRESSORS*

Miguel A. DELGADO
Universidad Carlos III de Madrid
Juan MORA
Universidad de Alicante

This paper deals with the practical situation of estimating nonparametric and semiparametric models when some or all regressors are discrete. When all regressors are discrete, Delgado and Mora (1995) show that a naive non-smoothing estimate produces globally consistent estimates of the regression function. Here we discuss that, in certain circumstances, it is advisable to use a smooth estimate. We show that the most popular smoothers, like kernels or k-NN, are asymptotically equivalent to the non-smoothing estimate. We also consider the mixed case where some regressors are discrete and others are continuous. The nonparametric estimates we present are useful in semiparametric problems. We discuss in detail the partially linear regression model and shape-invariant modelling. The paper also includes a Monte Carlo study.

1. Introduction

In econometric practice, few explanatory variables in regression models are continuous. Many of them are dummies, qualitative variables or counts; and others, though continuous in nature, are recorded at intervals and can be treated as discrete. This paper is concerned with the practical situation of estimating nonparametric and semiparametric models when some or all regressors are discrete.

In Section 2 we show global consistency and root-n-consistency results for several smoothing procedures in situations where all or some regressors are discrete. When all regressors are discrete, the simplest estimate consists of an average of all observations of the dependent variable with the same regressor value. Delgado and Mora (1995) show global consistency (in the sense of Stone, 1977; see Section 2.2 below) of this non-smoothing estimate

* This article is based on research funded by Spanish Dirección General de Investigación Científica y Técnica (DGICYT), reference number PB92-0247. We are grateful to two anonymous referees for their comments.

and prove its asymptotic equivalence (in a sense which is specified below), with respect to k-nearest neighbours. Here we show asymptotic equivalence with respect to other smoothing procedures like the regressogram and the popular kernel estimates. In Section 3 we provide asymptotic normality results for semiparametric estimates in models where root-n-consistency is not easy to achieve due to bias terms. We discuss in detail the semiparametric partly linear regression model and shape-invariant modelling with discrete or mixed regressors. In Section 4 we provide Monte Carlo simulations which form a basis for discussion on when smoothing is advisable in the presence of discrete regressors. Proofs are confined to an appendix.

2. Nonparametric weights with discrete and mixed regressors

Let Z be an $\mathbb{R}^q$ –valued discrete random variable. That is,

$$\exists\, \mathcal{D} \subset \mathbb{R}^q, \mathcal{D} \text{ countable set, with } P(Z \subset \mathcal{D}) = 1 \text{ and } \wp_i \in \mathcal{D} \Rightarrow P(Z = \wp_i) > 0. \quad [1]$$

Let (ζ_1, Z_1), ..., (ζ_n, Z_n) be independent and identically distributed (i.i.d.) random vectors. In this section we present asymptotic properties of alternative nonparametric estimates of the conditional expectation (or *regression function*) $m_\zeta(\wp) \equiv E[\zeta | Z = \wp]$.

2.1. *Nonparametric regression estimates with discrete regressors*

When regressors are discrete, $m_\zeta(\wp)$ can be estimated by

$$\hat{m}_\zeta(\wp) = \sum_j \zeta_j W_{nj}(\wp),$$

where, hereafter, summations run from 1 to n unless otherwise stated, and the nonparametric weights are defined as

$$W_{nj}(\wp) = I(Z_j = \wp) / (\sum_k I(Z_k = \wp)),$$

where $I(A)$ is the indicator function of event A and, hereafter, we arbitrarily define $0/0$ to be 0. Observe that these nonparametric weights do not require any smoothing value and, hence, we will refer to $\hat{m}_\zeta(\wp)$ as the *non-smoothing* estimate. When the sample size is small and there are many different values of Z in the sample, it may be convenient to smooth. We will consider three popular nonparametric smoothing estimates of the regression function, the regressogram, kernels and k-nearest neighbours.

Regressogram weights are defined as

$$\overset{\circ}{W}_{nj}(\wp) = I(\mathcal{B}_{nj}(\wp)) / (\sum_k I(\mathcal{B}_{nk}(\wp))),$$

where $\mathcal{B}_{nj}(\wp) = \{\exists\, i,\ 1 \le i \le k(n) : \wp \in J_i,\ Z_j \in J_i\}$ and J_1, ..., $J_{k(n)}$ are pairwise disjoint subsets such that $\cup_{j=1}^{k(n)} J_j = \mathbb{R}^q$. The corresponding *regressogram* estimate of $m_\zeta(\wp)$ is

$$\overset{\circ}{m}_\zeta(\wp) = \sum_j \zeta_j \overset{\circ}{W}_{nj}(\wp).$$

Usually, $J_1, \ldots, J_{k(n)}$ are adjacent subsets and then the regressogram looks like the popular histogram; this is precisely the reason why this estimate is called regressogram. When studying its asymptotic properties, we have to assume that

$$V_n \equiv \max_{1 \le i \le k(n)} V(J_i) \to 0, \text{ (as } n \to \infty), \qquad [2]$$

where $V(S)$ denotes the volume of the set S (using the standard measure in $\mathbb{R}^q$). The main advantage of these weights is that they are easy to compute.

Kernel weights are defined as

$$\tilde{W}_{nj}(\wp) = \psi((\wp - Z_j) / h_n) / \Sigma_k \psi((\wp - Z_k) / h_n),$$

where ψ is a function from $\mathbb{R}^q$ to $\mathbb{R}$ and h_n is a sequence of positive real numbers. The *kernel* estimate of $m_\zeta(\wp)$ is

$$\tilde{m}_\zeta(\wp) = \Sigma_j \zeta_j \tilde{W}_{nj}(\wp).$$

This estimate, which was first defined by Nadaraya (1964) and Watson (1964), is the most popular one in the nonparametric literature. We will assume that

$$\exists\, M > 0 \text{ such that } \|x\| \ge M \Rightarrow \psi(x) = 0 \text{ and } h_n \to 0 \text{ (as } n \to \infty). \qquad [3]$$

The *k-nearest neighbour estimate* of $m_\zeta(\wp)$ (hereafter referred to as k-NN estimate) is defined as follows: let $Z^{(j)}$ be the jth coordinate of Z $(1 \le j \le q)$, and s_{nj} the sample standard deviation of $Z_1^{(j)}, \ldots, Z_n^{(j)}$. First of all, we define for $u, v \in \mathbb{R}^q$

$$\rho_n(u, v) = (\Sigma_j ((u^{(j)} - v^{(j)}) / s_{nj})^2)^{1/2},$$

where the sum extends over all j, $1 \le j \le q$, such that $s_{nj} > 0$. Let c_{in} $(1 \le i \le n)$ be constants satisfying

$$\Sigma_i c_{in} = 1, \; c_{1n} \ge \ldots \ge c_{nn} \ge 0.$$

Define now for a given i $(1 \le i \le n)$

$$e(i, n, \wp) \equiv \#\{j : 1 \le j \le n, \rho_n(Z_j, \wp) = \rho_n(Z_i, \wp)\},$$
$$d(i, n, \wp) \equiv \#\{j : 1 \le j \le n, \rho_n(Z_j, \wp) < \rho_n(Z_i, \wp)\}.$$

A sequence of nonparametric weights can then be defined as

$$\omega_{ni}(\wp) = (\Sigma_{k=1}^{e(i, n, \wp)} c_{d(i, n, \wp) + k}) / e(i, n, \wp).$$

And the corresponding nonparametric estimate of $m_\zeta(\wp)$ is

$$\breve{m}_\zeta(\wp) = \Sigma_j \zeta_j \omega_{nj}(\wp).$$

Given a sequence k_n, the nonparametric estimate $\breve{m}_\zeta(\wp)$ is said to be a k-NN estimate if the following condition holds,

$$i > k_n \Rightarrow c_{in} = 0.$$

There are different possible k-NN estimates, according to various choices of the sequence c_{in}. Some possible c_{in} are defined in Stone (1977) (see also Devroye, 1978). The uniform k-NN estimate ($c_{in} = I(1 \leq i \leq k_n)/k_n$) is, possibly, the most popular one. In this case

$$\omega_{ni}(\wp) = \begin{Bmatrix} 1/k_n & \text{if } \rho_n(Z_i, \wp) < \rho_{nk}(\wp) \\ (k_n - d_{nk}(\wp))/(k_n\, e_{nk}(\wp)) & \text{if } \rho_n(Z_i, \wp) = \rho_{nk}(\wp) \\ 0 & \text{if } \rho_n(Z_i, \wp) > \rho_{nk}(\wp) \end{Bmatrix},$$

where now $\rho_{nk}(\wp)$ is the *k-th* value obtained after sorting the sequence of values $\rho_n(Z_1, \wp)$, ..., $\rho_n(Z_n, \wp)$, and $d_{nk}(\wp)$, $e_{nk}(\wp)$ are

$$e_{nk}(\wp) \equiv \#\{j : 1 \leq j \leq n,\ \rho_n(Z_j, \wp) = \rho_{nk}(\wp)\},$$

$$d_{nk}(\wp) \equiv \#\{j : 1 \leq j \leq n,\ \rho_n(Z_j, \wp) < \rho_{nk}(\wp)\}.$$

The k-NN weights are intuitively appealing. All nonparametric regression estimates can be viewed as local averages around the point at which regression is evaluated; with the k-NN estimates, one decides how many points are used in these local averages.

2.2. *Global consistency*

When regressors are discrete, the non-smoothing weights satisfy a property of global consistency, as Delgado and Mora (1995) prove. This means that $E\|\hat{m}_\zeta(Z) - m_\zeta(Z)\|^r = o(1)$ whenever the non-smoothing estimate is constructed using i.i.d. observations and $E\|\zeta\|^r < \infty$. Using this result we prove in this Section global consistency of the smoothing weights defined in Section 2.1.

We will first analyse the difference between non-smoothing and kernel weights. Let us assume that

$$\exists\, t \geq 0 \text{ such that } n^t [1 - p_n(\wp)^n] \to 0 \text{ (as } n \to \infty) \text{ uniformly in } \wp, \qquad [4]$$

where $p_n(\wp) \equiv 1 - P(0 < \|Z - \wp\| < h_n M)$, for M as defined in [3]. Then,

Theorem 1: *If* [1], [3] *and* [4] *hold,* $E\|\zeta\| < \infty$ *and* (ζ, Z), (ζ_1, Z_1), ..., (ζ_n, Z_n) *are i.i.d. random variables then* $P\{\hat{m}_\zeta(Z) \neq \tilde{m}_\zeta(Z)\} = o(n^{-t})$ *for* t *satisfying* [3].

Note that assumptions in Theorem 1 do not exclude discrete variables whose support contains accumulation points, but [4] restricts the probability mass which can be contained in the neighbourhoods of any accumulation point. If [4] does not hold, then it may happen that $P\{\hat{m}_\zeta(Z) \neq \tilde{m}_\zeta(Z)\}$ does not converge to *0*, as in the following example.

Example 1: Consider the discrete random variable Z with probability function $P(Z = 1) = 1/2$ *and* $P(Z = 1-j^{-1}) = c_0/j^2$ *for* $j = 1,2,...$, *where* $c_0 \equiv 3/\pi^2$. *Let* ζ *be any real random variable with* $E\|\zeta\| < \infty$ *satisfying*

$$\exists\, c_1 \in \mathbb{R} \text{ such that } P(\zeta \geq c_1 \mid Z=1) = 1 \text{ and } \forall\, \wp \in \mathcal{D}, \wp \neq 1 \Rightarrow P(\zeta < c_1 \mid Z = \wp) = 1. \quad [5]$$

For any symmetric function $\psi(.)$ *with support* $(-1,1)$, *if* $h_n = n^{-\gamma}$ *for any real number* $\gamma \in (0,1)$ *then,* $P\{\hat{m}_\zeta(Z) \neq \tilde{m}_\zeta(Z)\}$ *does not converge* 0, *as we prove in the appendix. Condition* [5] *ensures that* $m_\zeta(.)$ *is not a constant function in any neighbourhood of 1.*

If we compare Theorem 1 with Theorem 3 in Delgado and Mora (1995) (where the relationship between non-smoothing and *k*-NN estimates is analysed) we observe that no similar assumption to [4] is required when using *k*-NN weights. The reason why this happens is because if $\wp$ is an accumulation point in $\mathcal{D}$, asymptotically, all points which will be used to compute its *k*-NN estimate (that is, the *k* nearest neighbours of $\wp$) will satisfy $Z_j = \wp$; but some points which do not satisfy this condition will be used to compute the kernel estimate and their influence cannot be ignored.

The following assumption, stronger than [4], is much more intuitive:

$$\exists\, \mu > 0 \text{ such that } \forall\, \wp_1, \wp_2, \in \mathcal{D}, \wp_1 \neq \wp_2 \Rightarrow \|\wp_1 - \wp_2\| \geq \mu. \quad [6]$$

All usual discrete random variables (e.g. Poisson, negative binomial, and geometric) satisfy [6], and it is easily checked that [3] and [6] imply [4] (and, therefore, Theorem 1) for all $t \in \mathbb{R}$ (observe that, in that case, for *n* large enough $p_n(\wp) = 1\ \forall\, \wp \in \mathcal{D}$).

Both kernel weights and regressograms satisfy the following property of global consistency.

Theorem 2: Assume that [1] *and* [6] *hold,* $E\|\zeta\|^r < \infty$ *and* (ζ, Z), (ζ_1, Z_1), ..., (ζ_n, Z_n) *are i.i.d. random vectors.*

a) If [3] *holds, then* $E\|\tilde{m}_\zeta(Z) - m_\zeta(Z)\|^r = o(1)$.

b) If [2] *holds, then* $E\|\overset{o}{m}_\zeta(Z) - m_\zeta(Z)\|^r = o(1)$.

Devroye and Wagner (1980) proved a similar result to Theorem 2a considering jointly discrete and continuous regressors and under somewhat stronger conditions than [3]. Devroye and Wagner need conditions on the kernel function which exclude, among others, higher order kernels. They also need conditions on nh_n^q.

As for k-NN weights, if the following condition holds,

$$1/k_n + k_n/n \to 0 \text{ (as } n \to \infty), \qquad [7]$$

then, applying Stone's (1977) results, we know that the k-NN estimates satisfy a global consistency result similar to Theorem 2. On the other hand, Delgado and Mora (1995) prove an asymptotic equivalence result between k-NN and non-smoothing weights.

2.3. *Pointwise root-n-consistency*

When regressors are discrete, all nonparametric estimates defined above are root-n-consistent. We assume that ζ is an $\mathbb{R}^s$-valued random variable satisfying

$$E[\zeta\zeta'] < \infty \qquad [8]$$

Given $\wp \in \mathcal{D}$, denote $p(\wp) \equiv P(Z = \wp)$, $\Sigma(\wp) \equiv Var(\zeta \mid Z = \wp)$ and $\Gamma(\wp) \equiv p(\wp)^{-1} \Sigma(\wp)$, which can be estimated by $\hat{p}(\wp)$, $\hat{\Sigma}(\wp)$ and $\hat{\Gamma}(\wp)$ respectively, defined as

$$\hat{p}(\wp) = n^{-1}\Sigma_j I(Z_j = \wp),$$

$$\hat{\Sigma}(\wp) = \Sigma_j \zeta_j \zeta_j' W_{nj}(\wp) - \hat{m}_\zeta(\wp)\, \hat{m}_\zeta(\wp)'$$

$$\hat{\Gamma}(z) = \hat{p}(\wp)^{-1} \hat{\Sigma}(\wp).$$

Given $\wp_1, ..., \wp_f$, let $\Gamma(\wp_1, ..., \wp_f)$ and $\hat{\Gamma}(\wp_1, ..., \wp_f)$ be block diagonal $sf \times sf$ matrices with components $\Gamma(\wp_j)$ and $\hat{\Gamma}(\wp_j)$ $(1 \le j \le f)$ respectively. Then, we have

Theorem 3: If [1] *and* [8] *hold,* $(\zeta_1, Z_1), ..., (\zeta_n, Z_n)$ *are i.i.d. random vectors and* $\wp_j \in \mathcal{D}$ $(j = 1, ..., f)$ then

$$n^{1/2} \begin{pmatrix} \hat{m}_\zeta(\wp_1) - m_\zeta(\wp_1) \\ \cdots\cdots\cdots\cdots \\ \hat{m}_\zeta(\wp_f) - m_\zeta(\wp_f) \end{pmatrix} \xrightarrow{d} N(O, \Gamma(\wp_1, ..., \wp_f)),$$

and $\hat{\Gamma}(\wp_1, ..., \wp_f) \xrightarrow{p} \Gamma(\wp_1, ..., \wp_f)$.

Using Theorem 3 and the asymptotic equivalence results stated above we obtain,

Corollary 1: Assume that [1] *and* [8] *hold,* $(\zeta_1, Z_1), ..., (\zeta_n, Z_n)$ *are i.i.d. random vectors and* $\wp_j \in \mathcal{D}$ $(j = 1, ..., f)$. *If* [2] *and* [6] *hold, then*

$$n^{1/2} \begin{pmatrix} \mathring{m}_\zeta(\wp_1) - m_\zeta(\wp_1) \\ \cdots\cdots\cdots\cdots \\ \mathring{m}_\zeta(\wp_f) - m_\zeta(\wp_f) \end{pmatrix} \xrightarrow{d} N(O, \Gamma(\wp_1, ..., \wp_f)).$$

A similar result holds for the kernel estimate $\tilde{m}_\zeta(.)$ *if* [3] *and* [6] *hold, and for a* k-NN *estimate* $\breve{m}_\zeta(.)$ *if* [7] *holds.*

Of course, $\Gamma(\wp_1, ..., \wp_f)$ in Corollary 1 can be consistently estimated as in Theorem 3 or using smooth estimates.

A similar result to Corollary 1 for kernels was established by Bierens (1987) under different conditions.

All theorems and corollaries stated in this section will be used in Section 3 to prove asymptotic results in various semiparametric estimation problems.

2.4. Nonparametric regression estimates with mixed regressors

Let us suppose now that Z is a mixed regressor, that is, a random vector which contains both discrete and continuous variables or, more specifically,

$$\left.\begin{array}{l} Z = (Z^{(1)}, Z^{(2)}), \text{ where } Z^{(1)} \subset \mathbb{R}^r \text{ is discrete and} \\ Z^{(2)} \subset \mathbb{R}^s \text{ is absolutely continuous; } r + s = q,\ r \geq 1,\ s \geq 1. \end{array}\right\} \qquad [9]$$

We estimate $m_\zeta(\wp) \equiv E\,[\zeta \mid Z = \wp]$ using Nadaraya-Watson kernel weights (Nadaraya, 1964; Watson, 1964) for the continuous regressors and the non-smoothing weights for the discrete regressors, i.e.

$$W^*_{nj}(\wp) = \psi((\wp^{(2)} - Z^{(2)}_j)/h_n)\, I(\wp^{(1)} = Z^{(1)}_j) / \sum_k \psi((\wp^{(2)} - Z^{(2)}_j)/h_n)\, I(\wp^{(1)} = Z^{(1)}_k),$$

where ψ is a function from $\mathbb{R}^s$ to $\mathbb{R}$ defined as $\psi(z) = k(z_1)\, k(z_2) \ldots k(z_s)$, k is a function from $\mathbb{R}$ to $\mathbb{R}$ («kernel function») and h_n is a sequence of positive real numbers («smoothing values»). We estimate $m_\zeta(\wp)$ by

$$m^*_\zeta(\wp) = \sum_j \zeta_j W^*_{nj}(\wp),$$

for any random variable ζ. The asymptotic properties of these mixed weights cannot be derived using the same arguments as in previous sections, due to the presence of continuous variables. In fact, their behaviour is similar to that of Nadaraya-Watson weights with only q continuous regressors, provided that conditional density functions uniformly satisfy all standard assumptions in nonparametric estimation with absolutely continuous regressors (see Bierens, 1987, Section 3.2.).

3. Estimating semiparametric models with discrete regressors

Discrete regressors with possibly unbounded support are not a problem in some semiparametric models in which the focus of interest is to improve efficiency of the estimates. Stone's (1977) results with k-NN weights, allowing for very general regressors, were first applied by Robinson (1987) in semiparametric estimation in order to achieve asymptotic efficiency in

regression models in the presence of heteroskedasticity of unknown form (the same result had been obtained by Carroll, 1982, using kernels and under much more restrictive conditions on the regressors). These weights have been also applied to other semiparametric estimation problems by Newey (1990), Delgado (1992) and Delgado and Stengos (1994).

In many semiparametric inference problems, however, a bias term, which increases with the dimension of the regressors set, makes it difficult to achieve root-n-consistency results when using kernel weights. Robinson (1988) introduced the use of higher order kernels as a bias reduction technique in semiparametric problems. This technique has been also applied to other semiparametric procedures, like the average derivative method (Powell et al., 1989; Härdle and Stoker, 1989) and shape-invariant modelling (Pinkse and Robinson, 1995), among others.

When regressors are discrete, if non-smoothing weights are used then the bias term exactly equals *0*. Thanks to this property, the bias term which appears when kernel weights are used may be easily handled when regressors are discrete. In this section we discuss how this fact can be exploited to obtain asymptotic properties in semiparametric models with discrete regressors. We analyse in detail the partially linear regression model and shape-invariant modelling. We also make some remarks about how the same procedure may be used in other semiparametric estimation problems.

As expected, in the mixed case stronger conditions have to be imposed on the continuous part but no new techniques are required and theorems can be proved by combining the arguments in Section 2 with the well-developed asymptotic theory for continuous variables. We only analyse the mixed case in the partially linear model.

3.1. Partially linear regression model

Let (Y, X, Z) be an $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ -valued observable random variable such that

$$E[Y|X,Z] = \beta'X + \theta(Z) \quad a.s., \tag{10}$$

where β is an $\mathbb{R}^p$ -valued unknown parameter vector and θ is an unknown real function. Given a random sample $\{(Y_i, X_i, Z_i), i = 1, \ldots n\}$ from (Y, X, Z), if we define $\varepsilon_{\zeta i} \equiv \zeta_i - m_{\zeta i}$, where $m_{\zeta i} \equiv E[\zeta_i | Z_i]$, then,

$$\varepsilon_{Yi} = \beta'\varepsilon_{Xi} + U_i, \quad i = 1, 2, \ldots, n,$$

where $U_i = Y_i - E[Y_i | X_i, Z_i]$. Suppose that the following conditions hold,

$$E[U_i^2 | X_i, Z_i] = E[U_1^2] = \sigma^2 < \infty, \tag{11}$$

$$\Phi \equiv E[\varepsilon_{Xi}\, \varepsilon_{Xi}'] \text{ is positive definite (p.d.).} \tag{12}$$

Let us define $\bar{\Phi} = n^{-1}\sum_i \varepsilon_{Xi}\,\varepsilon'_{Xi}$ and the unfeasible estimate $\bar{\beta} = \bar{\Phi}^{-1}n^{-1}\sum_i \varepsilon_{Xi}\,\varepsilon_{Yi}$. Under [10], [11] and [12], $\bar{\beta}$ is asymptotically normal with

$$\text{AsyVar}\,(n^{1/2}\,(\bar{\beta}-\beta)) = \sigma^2\Phi^{-1}. \qquad [13]$$

Chamberlain (1992) has shown that [13] is a semiparametric asymptotic bound for model [10] in the absence of heteroskedasticity. Heckman (1986) and Engle et al. (1986) proposed feasible estimates of β using splines, but Rice (1986) proved that the rate of convergence for these estimates is slower than $n^{-1/2}$. Chen (1988) proposed an estimate of β based on a piecewise polynomial estimator of the unknown function θ, whereas Chen and Shiau (1991) proposed a two-stage spline smoothing estimate of β. They both proved that with those estimators root-n-consistency is achieved. Speckman (1988) and Robinson (1988, 1993) proposed feasible estimates of β by estimating the conditional expectations in ε_{Yi} and ε_{Xi}. We follow here this approach[1].

First we assume that Z is a discrete random variable. Given (ζ_i, Z_i), (ζ_1, Z_1), ..., (ζ_{i-1}, Z_{i-1}), (ζ_{i+1}, Z_{i+1}), ..., (ζ_n, Z_n) i.i.d. random vectors, $m_{\zeta i}(Z_i) \equiv E[\zeta_i|\ Z_i]$ is estimated by,

$$\tilde{m}_{\zeta i} = \sum_{j\neq i} \zeta_j\, \tilde{W}_{nj}\,(Z_i),$$

where now, for $i \neq j$

$$\tilde{W}_{nj}\,(Z_1) = \psi((Z_i - Z_j)/h_n)/\sum_{k\neq i}\psi((Z_i - Z_k)/h_n). \qquad [14]$$

Note that this is a «leave-one-out» estimate because ζ_i is not used to estimate $E[\zeta_i|\ Z_i]$. We use this estimate instead of an ordinary one in order to apply straightforwardly the global consistency results obtained in Section 2. Using these estimated residuals for $\zeta_i = Y_i, X_i$, it is possible to construct feasible estimates for Φ, β and σ^2. However, it is necessary to make a previous trimming in order to remove those observations for which the corresponding nonparametric estimate will not be accurate. Define the random variable

$$I_i = I(\sum_{k\neq i} I(Z_k = Z_i) > 0).$$

We can now construct $\tilde{\Phi} = n^{-1}\sum_i \tilde{\varepsilon}_{Xi}\tilde{\varepsilon}'_{Xi}I_i$, $\tilde{\beta} = \tilde{\Phi}^{-1}n^{-1}\sum_i \tilde{\varepsilon}_{Xi}\tilde{\varepsilon}_{Yi}I_i$ and $\tilde{\sigma}^2 = n^{-1}\sum_i (\tilde{\varepsilon}_{Yi} - \tilde{\beta}'\tilde{\varepsilon}_{Xi})^2 I_i$. The estimate $\tilde{\beta}$ achieves the semiparametric bound [13] under certain regularity conditions as stated in Theorem 4. If the support of Z contains accumulation points, stronger conditions on the kernel function have to be imposed; this is the reason why Theorem 4 contains two different assertions.

Theorem 4: Suppose [1], [3], [10], [11] and [12] *hold,* $E[U] < \infty$, $E[\theta(Z)^2] < \infty$, $E||X||^4 < \infty$ *and* (Y_1, X_1, Z_1), ..., (Y_n, X_n, Z_n) *are i.i.d. random vectors.*

[1] All notation used earlier will be redefined now in order to adapt it to the new assumptions.

a) If [6] *holds, then*

$$n^{1/2} (\tilde{\sigma}^2 \tilde{\Phi}^{-1})^{-1/2} (\tilde{\beta} - \beta) \xrightarrow{d} N(0, I_p).$$

b) If [4] *holds for some* $t \geq 1$ and ψ *is a bounded nonnegative function such that* $\forall\, x \in \mathbb{R}^q$, $\psi(x) \geq \mu I (\|x\| \geq \rho)$ *for some positive real constants* μ and ρ, then

$$n^{1/2} (\tilde{\sigma}^2 \tilde{\Phi}^{-1})^{-1/2} (\tilde{\beta} - \beta) \xrightarrow{d} N(0, I_p).$$

Note that, unlike Robinson (1988), it is not necessary to assume independence between regressors and regression errors. When we use non-smoothing or k-NN weights a similar result to Theorem 4 also holds (see Delgado and Mora 1995).

The homoskedasticity assumption can be easily removed but the asymptotic variance will change in the usual way (see e.g. Eicker 1963 and White 1980). Let us assume that, instead of [11], we have

$$E[U^2 | X, Z] = \sigma^2 (X, Z) > 0 \ a.s. \qquad [15]$$

When the support of Z contains no accumulation points, the result for the heteroskedastic model is as follows.

Corollary 2: If [1], [3], [6], [10], [12] and [15] *hold,* $E[U^4] < \infty$, $E||X||^4 < \infty$ *and* $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ *are i.i.d. random vectors, then*

$$n^{1/2} \tilde{\Psi}^{-1/2} (\tilde{\beta} - \beta) \xrightarrow{d} N(0, I_k),$$

where the matrix $\tilde{\Psi}$ *is defined by*

$$\tilde{\Psi} = \tilde{\Phi}^{-1} \{n^{-1} \Sigma_i (\tilde{\varepsilon}_{Yi} - \tilde{\beta}' \tilde{\varepsilon}_{Xi})^2 \tilde{\varepsilon}_{Xi} \tilde{\varepsilon}'_{Xi} I_i\} \tilde{\Phi}^{-1}.$$

Up to now we have only analysed the case when all regressors in the nonparametric part of the model are discrete. When there are both discrete and continuous regressors, the methodology is similar to that used when only continuous regressors are present, though notation and proofs become more lengthy and less intuitive. As in Section 2.4., we now estimate $m_{\zeta i} \equiv E[\zeta_i | Z_i]$ using Nadaraya-Watson kernel weights (Nadaraya, 1964; Watson, 1964) for the continuous regressors and the non-smoothing weights for the discrete regressors, i.e., using

$$W^*_{nj}(Z_i) = \psi_{ij}(h_n) I(Z_i^{(1)} = Z_j^{(1)}) / \Sigma_k \psi_{ik}(h_n) I(Z_i^{(1)} = Z_k^{(1)}), \qquad [16]$$

where hereafter we denote

$$\psi_{ij}(h_n) \equiv \psi((Z_i^{(2)} - Z_j^{(2)})/h_n),$$

and ψ and h_n are as in Section 2.4. We estimate $m_{\zeta i}$ by

$$m^*_{\zeta i} = \Sigma_j \zeta_j W^*_{nj}(Z_i),$$

for any random variable ζ. (Note that this is not a «leave-one-out» estimator). As in the discrete case, it is possible to construct estimated residuals $\varepsilon^*_{\zeta i}$ and estimates of the parameters of interest Φ^*, β^* and σ^{*2}, but now the trimming function is

$$I^*_i = I\,(\Sigma_k \psi_{ik}\,(h_n)\, I(Z^{(1)}_i = Z^{(1)}_k)/n h^q_n > b_n), \qquad [17]$$

where b_n is a sequence of positive real numbers (trimming values).

Some additional assumptions are required to prove that a similar result to Theorem 4 holds when there are both continuous and discrete regressors in the unknown part of the model, i.e., when Z satisfies [9]. First, following Robinson (1988), we define three classes of functions.

Given $\mu > 0$, the class G^{∞}_{μ} comprises all functions g: $\mathbb{R}^s \to \mathbb{R}$ satisfying:

i) $g(.)$ is $(m-1)$-times partially differentiable for $m - 1 < \mu \leq m$ and all z.

ii) $\exists\ \rho > 0$, $c \in \mathbb{R}$ such that for all z, if y is such that $\|y - z\| < \rho$ then $|R(y, z)| \leq c\|y - z\|^{\mu}$, where $R(y, z)$ is the remainder of a Taylor expansion of order $m - 1$ of $g(.)$ at z.

iii) $g(.)$ and its partial derivatives of order $m - 1$ and less are all bounded.

Given $\alpha > 0$, $\mu > 0$ and an absolutely continuous random variable Z with density function $f(.)$, the class $G^{\alpha}_{\mu}(Z)$ comprises all functions g: $\mathbb{R}^s \to \mathbb{R}$ satisfying:

i) $g(.)$ is $(m-1)$-times partially differentiable for $m - 1 < \mu \leq m$ and all z.

ii) $\exists\ \rho > 0$, and h: $\mathbb{R}^s \to \mathbb{R}$ such that for all z, if y is such that $\|y - z\| < \rho$ then $|R(y, z)| \leq h(z)\|y - z\|^{\mu}$, where $R(y, z)$ is the remainder of a Taylor expansion of order $m - 1$ of $g(.)$ at z.

iii) $g(.)$ satisfies that $\int |g(z)|^{\alpha} f(z)\, dz < \infty$, and this inequality also holds for $h(.)$ and for all partial derivatives of $g(.)$ of order $m - 1$ and less.

Given $l \in \mathbb{N}$, the class K_l comprises all functions k: $\mathbb{R} \to \mathbb{R}$ satisfying:

i) $\int k(u)\, du = 1$, $\int u^j k(u)\, du = 0$ if $1 \leq j \leq l - 1$.

ii) $\exists\ \varepsilon > 0$, $c > 0$ such that $\forall u \in \mathbb{R}\ \ |k(u)|(1 + |u|^{l+1+\varepsilon}) < c$.

The functions in G^{α}_{μ} and G^{∞}_{μ} are thus expanded in a Taylor series with a local Lipschitz condition on the remainder. On the other hand, K_l contains higher order kernels of order l satisfying a slightly stronger tail condition than $\int |u^l k(u)|\, du < \infty$, which is the usual condition required in the higher-order kernel literature.

Now, given $\wp \in \mathcal{D} \subseteq \mathbb{R}^r$, consider the following functions: $f_\wp$: $\mathbb{R}^s \to \mathbb{R}$ denotes the density function of $Z^{(2)} | Z^{(1)} = \wp$; $\theta_\wp$: $\mathbb{R}^s \to \mathbb{R}$ is defined as $\theta_\wp(a) = \theta(\wp, a)$ for $a \in \mathbb{R}^s$ ($\theta(.,.)$ as in [10] for $(\wp, a) \in \mathbb{R}^q$); and $\xi_\wp$: $\mathbb{R}^s \to \mathbb{R}$ is defined as $\xi_\wp(a) = E[X | Z^{(1)} = \wp, Z^{(2)} = a]$ for $a \in \mathbb{R}^s$. We assume that

$$U \equiv Y - E[Y | X, Z] \text{ and } (X, Z) \text{ are independent,} \qquad [18]$$

$$\left.\begin{array}{c} \exists v \in \mathbf{N}: f_\wp \in G_v^\infty,\ \theta_\wp \in G_v^4 (Z^{(2)} | Z^{(1)} = \wp) \text{ and} \\ \xi_\wp \in G_v^2 (Z^{(2)} | Z^{(1)} = \wp) \text{ uniformly in } \mathcal{D} \end{array}\right\} \qquad [19]$$

$$b_n \to 0, \quad n b_n^{-4} h_n^{4v} \to 0, \quad n b_n^4 h_n^{2s} \to \infty \text{ (as } n \to \infty) \text{ and} \qquad [20]$$

$$\text{the kernel function } k \text{ is in class } K_{2v-1}. \qquad [21]$$

Uniformly in $\mathcal{D}$ means that the constants which appear in the definition do not depend on the value $\wp$. Assumption [19] specifies the degree of smoothness in $f_\wp$, $\theta_\wp$ and $\xi_\wp$ which is required. Assumption [20] gives conditions on the rate of convergence of h_n and b_n. Assumption [21] specifies the relationship between the degree of smoothness and the order of the kernel function. Observe that, as a consequence of [20], $2v > s$ and, hence, ψ is at least of order s. Note also that [21] does not imply that $k \in K_{2v}$ *as* ψ may not satisfy the tail condition required for functions in K_{2v}.

Theorem 5: If [9], [10], [11], [12], [18], [19], [20] *and* [21] *hold,* $E||X||^4 < \infty$ *and* $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ *are i.i.d. random vectors, then,*

$$n^{1/2} \sigma^{*-1} \Phi^{*1/2} (\beta^* - \beta) \xrightarrow{d} N(0, I_k).$$

Observe that the conditions required in Theorem 5 are much stronger than those required in Theorem 4 –this is why we have preferred to consider first the discrete case separately, as in many cases all variables in the unknown part of the model are discrete. Also observe that, unlike in Theorem 4, independence between regressors and regression errors is assumed in Theorem 5. Hence, this result does not follow straightforwardly in the heteroskedastic model.

As noted in previous sections, when the sample size is small and there are many different values of $Z^{(1)}$ in the sample, it may be necessary to smooth in the discrete part as well. In such a case Theorem 5 does not apply directly but, using the equivalence results stated in Section 2, it is easy to deduce from Theorem 5 similar results for estimates which utilise k-NN, kernel or regressogram weights in the discrete part of the model.

3.2. Shape-invariant modelling

Let us assume that (ζ, Z), (ζ^*, Z^*) are both $\mathbb{R} \times \mathbb{R}^q$-valued observable random variables such that Z and Z^* are discrete that is,

$\exists\ \mathcal{D} \subset \mathbb{R}^q$, $\mathcal{D}$ countable set, such that $P(Z \subset \mathcal{D}) = 1$, $P(Z^* \subset \mathcal{D}) = 1$. [22]

We will denote F as the following subset of $\mathcal{D}$:

$$F \equiv \{\wp \in \mathcal{D} : P(Z = \wp) > 0 \text{ and } P(Z^* = \wp) > 0\}.$$

Note that we do not require that the probability function of Z and Z^* is positive in exactly the same points, but an assumption on F will be necessary –see [27] below.

Let us suppose that there exists a linear relationship between the regression functions $m(\wp) \equiv E[\zeta \mid Z = \wp]$ and $m^*(\wp) \equiv E[\zeta^* \mid Z^* = \wp]$, that is,

$$\exists\ \theta_0 = (\theta_{10}, \theta_{20}) \in \mathbb{R}^2\ (\theta_{20} \neq 0) \text{ such that } m^*(\wp) = \theta_{10} + \theta_{20}\, m(\wp)\ \forall \wp \in F \qquad [23]$$

Given independent random samples $\{(\zeta_i, Z_i),\ i = 1, .., n)\}$ and $\{(\zeta_j^*, Z_j^*), j = 1, .., n\}$[2], the objective of this section is to propose root-n-consistent estimates of the unknown parameter θ_0. We also discuss how our results may be extended to non-linear semiparametric relationships when regressors are discrete.

The relationship specified in equation [23] appears when $m(\wp)$ and $m^*(\wp)$ are functions with similar shape, but there is no reasonable parametric model for each regression function. Parameter θ_{10} is related to changes in location, whereas parameter θ_{20} is related to changes in scale. Lawton et al. (1972) and Gasser et al. (1984) (among others) provide with examples in which similar models to [23] may apply. In econometric practice, these models are likely to appear when analysing certain microeconomic data. Consider, for instance, the case in which (ζ, Z) are, respectively, «percentage of expenditure on food» and «age of the reference person» for households in a low level of income and (ζ^*, Z^*) are the same variables but considered for households in a high level of income. After a nonparametric analysis of data, it may seem unreasonable to assume that $m(\wp)$ and $m^*(\wp)$ are the same function; but it may be possible that a relationship as [23] holds and then it will be of interest to estimate θ_0.

Some recent papers have analysed similar models to [23] in settings which are different from ours. Härdle and Marron (1990) consider a (possibly non-linear) parametric relationship between the two unknown regression functions when regressors are fixed and taken equally spaced on the unit interval. The θ_{20} estimate can be used for testing whether the two regression curves have equal shape. Pinkse and Robinson (1995) consider the same kind of relationship as Härdle and Marron (1990) when regressors are continuous random variables, and prove that a more efficient estimate is obtained by pooling the two data sets.

[2] We assume that the size of both random samples is the same for the sake of simplicity. This assumption is, obviously, not necessary.

The true parameter θ_0 satisfies that

$$\theta_0 = \underset{\theta}{argmin}\ Q(\theta) \equiv \underset{\theta}{argmin}\ \Sigma_{\wp \in F} (m^*(\wp) - \theta_1 - \theta_2\, m(\wp))^2\, v(\wp) \qquad [24]$$

($v(.)$ is any positive real «weight function», chosen in such a way that the summation is finite). We can obtain a feasible estimate replacing the unknown regression functions by non-smoothing estimates $\hat{m}(\wp)$ and $\hat{m}^*(\wp)$ defined as in Section 2. Thus, let us define the *least squares estimate.*

$$\hat{\theta} \equiv (\hat{\theta}_1, \hat{\theta}_2) = \underset{\theta}{argmin}\ \Sigma_{\wp \in F} (\hat{m}^*(\wp) - \theta_1 - \theta_2\, \hat{m}(\wp))^2\, \hat{w}_n(\wp).$$

where the weight function we consider here is

$$\hat{w}_n(\wp) = I(n^{-1}\Sigma_j I(Z_j = \wp) \geq \varepsilon) \times I(n^{-1}\Sigma_j I(Z_j^* = \wp) \geq \varepsilon),$$

for a fixed real value $\varepsilon > 0$. We assume that ε is taken in such a way that

$$\varepsilon \notin \{p \in (0,1): \exists\, \wp \in \mathcal{D} \text{ such that } P(Z = \wp) = p \text{ or } P(Z^* = \wp) = p\} \qquad [25]$$

This is a mere technical condition which does not restrict, in practice, the choice of ε. This condition is introduced in order to ensure that $\forall\, \wp \in F$, $\hat{w}_n(\wp)$ converges to $w(\wp)$, where we denote

$$w(\wp) = I(P(Z = \wp) > \varepsilon) \times I(P(Z^* = \wp) > \varepsilon).$$

The value ε must also satisfy condition [27] below, where we also discuss how this value can be chosen in practice.

We assume in our model that

$$E[\zeta^2] < \infty,\ E[\zeta^{*2}] < \infty, \qquad [26]$$

$\exists\, \wp_1, \wp_2 \in F$ such that:

$$\left.\begin{array}{l} a)\ m(\wp_1) \neq m(\wp_2) \\ b)\ P(Z = \wp_i) > \varepsilon,\ P(Z^* = \wp_i) > \varepsilon \text{ for } i = 1, 2, \end{array}\right\} \qquad [27]$$

$$\text{If } \wp \in F,\ Var(\zeta \mid Z = \wp) > 0,\ Var(\zeta^* \mid Z^* = \wp) > 0. \qquad [28]$$

Assumption [26] ensures that we can apply the asymptotic results proved in Section 2. Assumption [27] is an identifiability condition: it ensures that θ_0 is the only solution to [24] when $w(\wp)$ is used as weight function. Assumption [28] avoids degenerate cases which could be treated in a simpler way. Let us define

$$\lambda(\wp) = \Gamma^*(\wp) + \theta_{20}^2\, \Gamma(\wp),$$

$$\hat{\lambda}(\wp) = \hat{\Gamma}^*(\wp) + \hat{\theta}_2^2\, \hat{\Gamma}(\wp),$$

where $\Gamma(\wp)$, $\Gamma^*(\wp)$, $\hat{\Gamma}^*(\wp)$ and $\hat{\Gamma}(\wp)$ are as defined in Section 2. Then we have the following result.

Theorem 6: If [22], [23], [25], [26], [27], *and* [28] *hold and* (ζ_1, Z_1), (ζ_1^*, Z_1^*), ..., (ζ_n, Z_n), (ζ_n^*, Z_n^*), *are i.i.d. random vectors, then*

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A^{-1}VA^{-1}),$$

where the matrices A and V are defined by

$$A \equiv \sum_{\wp \in F} \begin{pmatrix} 1 & m(\wp) \\ m(\wp) & m(\wp)^2 \end{pmatrix} w(\wp),$$

$$V \equiv \sum_{\wp \in F} \begin{pmatrix} 1 & m(\wp) \\ m(\wp) & m(\wp)^2 \end{pmatrix} \lambda(\wp) w(\wp).$$

Furthermore, the asymptotic variance-covariance matrix can be consistently estimated by $\hat{A}^{-1}\hat{V}\hat{A}^{-1}$, *where* $\hat{A}$ *and* $\hat{V}$ *are defined as A, V, replacing* $m(\wp)$, $w(\wp)$ *and* $\lambda(\wp)$, *by* $\tilde{m}(\wp)$, $\hat{w}(\wp)$, *and* $\tilde{\lambda}(\wp)$.

According to the definition of $w(\wp)$, the summation in A and V extend only over a finite number of terms. Moreover, A is positive definite as a consequence of [27] and Cauchy inequality. The non-smoothing estimates used in this theorem may be replaced by kernel or k-NN estimates (a similar proof applies).

On implementing this estimate, the practitioner only has to choose the fixed value ε. If the asymptotic variance-covariance matrix were known, obviously ε should be such that the most efficient estimate were obtained. In practice, the choice of this value must depend on the sample size and variance of Z and Z^*, the objective of this choice being to consider only those points for which we have accurate estimates.

The asymptotic variance-covariance matrix of $\hat{\theta}$ reminds us of the «heteroskedastic» nature of the model. Observe that

$$\text{AsyVar}\,(n^{1/2}(\hat{m}^*(\wp) - \theta_{10} - \theta_{20}\,\hat{m}(\wp)) = \lambda\,(\wp).$$

As usual, we can obtain a more efficient estimate in a second stage if we use weighted least squares. Specifically, let us define the *generalised least squares estimate as*

$$\tilde{\theta} \equiv (\tilde{\theta}_1, \tilde{\theta}_2) = \underset{\theta}{argmin} \sum_{\wp \in F} (\hat{m}^*(\wp) - \theta_1 - \theta_2\,\hat{m}(\wp))^2\,\hat{\lambda}\,(\wp)^{-1}\,\hat{u}_n(\wp),$$

where the trimming function we consider now is

$$\hat{u}_n(\wp) = I(n^{-1}\textstyle\sum_j I(Z_j = \wp) \geq \rho/n^{\alpha}) \times I(n^{-1}\sum_j I(Z_j^* = \wp) \geq \rho/n^{\alpha}),$$

for fixed positive real values ρ and α. Observe that, unlike $\hat{w}_n(\wp)$, the trimming function $\hat{u}_n(\wp)$ satisfies that

$$\forall\ \wp \in F,\ \hat{u}_n(\wp) \xrightarrow{p} 1.$$

Hence, asymptotically all values in F are taking into account on computing $\hat{\theta}$ irrespective of the values ρ and α we choose. We assume that

$$\exists\ \delta > 0 \text{ such that } \forall\ \wp \in F \quad Var(\zeta \mid Z = \wp) > \delta \text{ and } Var(\zeta^* \mid Z^* = \wp) > \delta, \quad [29]$$

Theorem 7: If [22], [23], [25], [26], [27], *and* [29] *hold and* (ζ_1, Z_1), (ζ_1^*, Z_1^*), ..., (ζ_n, Z_n), (ζ_n^*, Z_n^*), *are i.i.d. random vectors, then*

$$n^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega^{-1}),$$

where the matrix Ω *is defined by*

$$\Omega \equiv \sum_{\wp \in F} \begin{pmatrix} 1 & m(\wp) \\ m(\wp) & m(\wp)^2 \end{pmatrix} \lambda(\wp)^{-1}.$$

Furthermore, Ω *may be consistently estimated by* $\hat{\Omega}$, *defined in the same way as* Ω *replacing* $m(\wp)$ *and* $\lambda(\wp)$, *by* $\hat{m}(\wp)$ *and* $\hat{\lambda}(\wp)$.

If we compare Theorems 6 and 7 we observe that there are at least two reasons why $\tilde{\theta}$ is preferable to $\hat{\theta}$: on the one hand, $\tilde{\theta}$ is more efficient than $\hat{\theta}$ (it is easy to prove that $A^{-1}VA^{-1} - \Omega^{-1}$ is positive definite); on the other hand, the asymptotic distribution of $n^{1/2}(\tilde{\theta} - \theta_0)$ does not depend on the choice of any real number, whereas the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta_0)$ may be severely affected by a bad choice of ε. In Section 4 we analyse the finite-sample behaviour of both estimates in various statistical models.

The linear relationship considered in [23] may be too simple to capture the true nature of the observations. More general parametric relationships may be considered. Specifically, $m^*(\wp) = S(\theta, m(\wp))$, where $S(.,.)$ is a known real function and θ is an unknown vector of parameters, may be a more realistic assumption than [23]. But, essentially, the same ideas which underlie our proposed estimate may be also used in this case –we prefer the simpler model [23] for the sake of clarity. Even more general models can be considered, such as $m^*(\wp) = S(\theta_1, m(T(\theta_2, \wp)))$ for known real functions $S(.,.)$ $T(.,.)$ and unknown vector parameters, θ_1, θ_2. But here the function T and the parameter space must be such that $T(\theta_2, \wp) \in \mathfrak{D}$. Hence, strong conditions should be imposed on $T(.,.)$, the parameter space and the estimates of θ_2.

3.3. Other semiparametric models

In some semiparametric problems it is not straightforward to achieve root-n-consistency owing to the bias introduced by the nonparametric estimate, as in the models studied in previous sections or in the «average derivative estimation (ADE) method» (see e.g. Powell et al. 1989, Härdle and Stoker 1989 or Robinson 1989). In the ADE model Chamberlain (1986) proved that if all regressors are discrete then the parameter of interest may not be identifiable (even up to a scale coefficient). In the mixed continuous-discrete case, it would be possible to achieve root-n-consistency, but the involved resulting model will probably not capture the true relationship between the variables concerned (see Stoker, 1991, Section 5.2.a.).

In other semiparametric problems, the goal is to improve efficiency rather than achieve root-n-consistency. In most of these models, implementation of discrete regressors using our methods is straightforward. For instance, in the asymptotic efficient estimation in the presence of heteroskedasticity of unknown form, Robinson (1987) proved (using k-NN regression estimates) that the semiparametric estimate is asymptotically efficient even when regressors have discrete or mixed distribution. As a consequence of our results in Section 2, when all regressors are discrete the same asymptotic distribution is obtained using non-smoothing, regressogram or kernel weights. Nonparametric k-NN weights have been also used in other semiparametric inference problems in which weights presented in this paper are also straightforwardly applicable (see e.g. Newey, 1990, and Delgado, 1992).

4. Simulations

We have generated observations from the regression models discussed in Sections 3.1. and 3.2. and computed the various semiparametric estimates discussed there. The results are contained in Tables 1, 2, 3, 4 and 5.

First we have generated observations from eight partially linear regression models. In models 1-6 we have taken X and Z to be scalar random variables. In these six models Z was taken from a Poisson distribution with mean λ (specified below) and X was taken as $X = Z + V$, where V was generated from a normal population independent of Z with zero mean and variance 1. In all models the error term U is independent from V and was generated from a normal population with zero mean and variance $\sigma^2(Z)$. The complete description of models 1-6 is as follows:

Model	λ	$\sigma_u^2(Z)$	Underlying model for Y
1	0.3	1	$Y = 1 + X + Z + U$
2	3.0	1	$Y = 1 + X + Z + U$
3	0.3	1	$Y = 1 + X - 3(Z-1)^2 + U$
4	3.0	1	$Y = 1 + X - 3(Z-1)^2 + U$
5	0.3	$(1+Z/3)^2$	$Y = 1 + X - 3(Z-1)^2 + U$
6	3.0	$(1+Z/3)^2$	$Y = 1 + X - 3(Z-1)^2 + U$

Note that models 1 and 2 are linear and homoskedastic, models 3 and 4 are nonlinear and homoskedastic and models 5 and 6 are nonlinear and heteroskedastic. In models with uneven label, the variance of Z is small and in every sample the majority of values will be 0 or 1; however, in models with even label samples will contain many different values of Z.

In models 7 and 8, Z was taken to be a bivariate Poisson distribution, $Z = (Z_1, Z_2)$ (both Z_1 and Z_2 with mean λ), V and U were as in models 1-6 and $X = Z_1 + Z_2 + V$. The complete description of these models is:

Model	λ	$\sigma_u^2(Z)$	Underlying model for Y
7	0.3	1	$Y = 1 + X - 3(Z_1 - 1)^2 - 3(Z_2 - 1)^2 + U$
8	3.0	1	$Y = 1 + X - 3(Z_1 - 1)^2 - 3(Z_2 - 1)^2 + U$

In all models the semiparametric estimates $\hat{\beta}$ $\tilde{\beta}$ and $\check{\beta}$ (kernel, non-smoothing and uniform k-NN estimates, respectively) were computed. In models 1-6 the kernel we used was the *Epanechnikov kernel* (the most efficient one in nonparametric estimation), defined as

$$k(u) = 0.75(1 - u^2) I(|u| \leq 1).$$

In models 7-8 the kernel used was the product of two univariate Epanechnikov kernels. On computing both the kernel and the k-NN estimates smoothing values (h_n and k_n respectively) have to be selected. We have simply selected three possible h_n and k_n trying to cover meaningful intervals for them. Observe that, according to our selection of the support $\mathcal{D}$ and the kernel function k, if $h_n < 1$ then the kernel estimate is the same as the non-smoothing one.

From the results in Section 3.1., the asymptotic distribution of the kernel estimate $\tilde{\beta}$ is

$$\text{Models 1-4, 7-8:} \quad n^{1/2}(\tilde{\beta} - 1) \xrightarrow{d} N(0,1),$$

$$\text{Model 5:} \quad n^{1/2}(\tilde{\beta} - 1) \xrightarrow{d} N(0,1.1^2).$$

$$\text{Models 6:} \quad n^{1/2}(\tilde{\beta} - 1) \xrightarrow{d} N(0,2^2).$$

The same asymptotic distributions hold for the non-smoothing and k-NN estimates.

We report the sample mean (M) and mean squared error (E) of each estimate. Table 1 contains results corresponding to a sample size of $n = 40$

observations; the reported values are based on $r = 2000$ replications. Tables 2 and 3 contain corresponding results for $n = 200$ and $n = 1000$, respectively.

TABLE 1
Sample size = 40, Number of replications = 2000
Non-Smoothing Estimate

	M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
M	1.0017	0.9995	1.0012	0.9993	1.0034	1.0006	1.0022	0.9993
E	0.0293	0.0381	0.0299	0.0381	0.0385	0.1667	0.0332	0.0925

Kernel Estimates ($h_1 = 1.25$, $h_2 = 1.75$, $h_3 = 2.25$)

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
h_1	M	1.0734	1.0595	1.1845	0.5008	1.1847	0.4998	1.3456	0.3296
	E	0.0311	0.0314	0.0735	0.4043	0.0735	0.4926	0.1635	0.8035
h_2	M	1.1174	1.0864	1.2946	0.1687	1.2995	0.1634	1.5165	−0.284
	E	0.0373	0.0345	0.1429	1.0132	0.1436	1.1064	0.3464	2.3846
h_3	M	1.1425	1.1742	1.3373	−0.776	1.3437	−0.887	1.5739	−1.265
	E	0.0454	0.0528	0.1664	3.8764	0.1612	4.0835	0.4145	6.5127

k-NN Estimates ($k_1 = 3$, $k_2 = 6$, $k_3 = 8$)

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
k_1	M	1.0256	1.0687	0.9023	−1.056	0.9032	−1.143	0.8536	−4.733
	E	0.0281	0.0413	0.0823	12.613	0.1040	14.123	0.1297	46.987
k_2	M	1.0314	1.1424	0.0834	−2.765	0.8834	−2.563	0.8663	−7.198
	E	0.0289	0.0584	0.0924	26.254	0.1045	25.632	0.1562	80.909
k_3	M	1.0356	1.1893	0.8791	−3.532	0.8821	−3.563	0.8953	−8.065
	E	0.0286	0.0683	0.0945	33.915	0.1149	34.264	0.1589	94.235

In nonparametric estimation, the typical trade-off between bias and variance is closely related with the degree of smoothing. Specifically, bias increases/decreases as the amount of smoothing increases/decreases and variance increases/decreases as the amount of smoothing decreases/increases. This behaviour is observed using any smoother. However, in semiparametric estimation problems this relationship is not so evident. In fact, we find in the simulations reported here that, for fixed sample size, the non-smoothing estimate can perform better than the others in terms of bias and variance, and this fact is stressed when the nonparametric part of the model exhibits high volatility.

TABLE 2
Sample size = 200, Number of replications = 2000
Non-Smoothing Estimate

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
	M	0.9998	0.9979	1.0006	0.9993	1.0008	0.9980	0.9991	1.0002
	E	0.0052	0.0052	0.0050	0.0052	0.0069	0.0250	0.0058	0.0081

Kernel Estimates ($h_1 = 1.15$, $h_2 = 1.45$, $h_3 = 1.75$)

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
h_1	M	1.0511	1.0284	1.0952	0.6958	1.0907	0.6973	1.1999	0.5076
	E	0.0075	0.0056	0.0152	0.1063	0.0155	0.1209	0.0466	0.2686
h_2	M	1.1022	1.0612	1.2130	0.3545	1.2077	0.3579	1.4037	−0.033
	E	0.0151	0.0083	0.0540	0.4461	0.0537	0.4590	0.1750	1.1400
h_3	M	1.1189	1.0775	1.2599	0.2147	1.2570	0.2129	1.4846	−0.249
	E	0.0188	0.0106	0.0780	0.6560	0.0774	0.6717	0.2486	1.6598

k-NN Estimates ($k_1 = 5$, $k_2 = 12$, $k_3 = 16$)

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
k_1	M	1.0046	1.0210	0.9603	0.2001	0.9636	0.1968	0.9012	−1.971
	E	0.0051	0.0059	0.0140	1.6441	0.0139	1.7434	0.0299	11.352
k_2	M	1.0146	1.0468	0.9236	−0.705	0.9196	−0.645	0.8384	−4.100
	E	0.0053	0.0080	0.0193	4.8134	0.0206	4.5358	0.0472	29.590
k_3	M	1.0187	1.0656	0.9027	−1.133	0.9045	−1.112	0.8207	−5.067
	E	0.0055	0.0105	0.0242	6.9217	0.0249	6.6985	0.0550	40.999

In models 1 and 2 (both linear) all estimates have similar behaviour; the *k-NN* estimates perform slightly better than the others in model 1 and the kernel estimates seem to be the best ones in model 2 (though, as expected, in both cases the non-smoothing estimate is the one with lowest bias). In models 3, 5 and 7 (nonlinear in Z and with low variance for Z) the non-smoothing estimate is the most adequate one, but the other nonparametric estimates also behave properly. In models 4, 6 and 8 (nonlinear in Z and with high variance for Z) the non-smoothing estimate is, again, the best one but, unlike in previous models, kernel and *k-NN* estimates perform rather poorly. In the heteroskedastic models the variance varies in the expected direction. In the two-dimensional models there is an increase in variance as a result of the poorer performance of the nonparametric estimate.

TABLE 3
Sample size = 1000, Number of replications = 2000
Non-Smoothing Estimate

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
	M	0.9986	1.0014	0.9997	0.9956	1.0102	1.0104	0.9891	1.0011
	E	0.0010	0.0011	0.0010	0.0011	0.0014	0.0043	0.0013	0.0012

Kernel Estimates ($h_1 = 1.05$, $h_2 = 1.15$, $h_3 = 1.25$)

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
h_1	M	1.0153	1.0062	1.0143	0.9315	1.0189	0.9198	1.0453	0.8551
	E	0.0012	0.0011	0.0014	0.0073	0.0016	0.0117	0.0032	0.0246
h_2	M	1.0509	1.0300	1.0923	0.7076	1.0867	0.7101	1.1803	0.4678
	E	0.0033	0.0019	0.0082	0.0887	0.0088	0.0891	0.0329	0.2882
h_3	M	1.0745	1.0467	1.1332	0.5446	1.1332	0.5446	1.2789	0.2021
	E	0.0066	0.0031	0.0194	0.2024	0.0204	0.2234	0.0787	0.6655

k-NN Estimates ($k_1 = 11$, $k_2 = 22$, $k_3 = 28$)

		M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8
k_1	M	1.0021	1.0078	0.9651	0.6897	0.9634	0.6563	0.9444	−0.612
	E	0.0009	0.0009	0.0024	0.1956	0.0024	0.2443	0.0064	2.9896
k_2	M	1.0024	1.0143	0.9624	0.3621	0.9610	0.3445	0.9216	−1.578
	E	0.0009	0.0010	0.0035	0.6124	0.0035	0.6342	0.0110	6.8954
k_3	M	1.0029	1.0174	0.9623	0.2280	0.9605	0.2170	0.9025	−1.924
	E	0.0009	0.0013	0.0035	0.7784	0.0037	0.8545	0.0121	9.5953

These results are not a surprise and can be explained in terms of the closeness between $m_\zeta(\wp_1)$ and $m_\zeta(\wp_2)$ when $\wp_1$ and $\wp_2$ are close values within $\mathcal{D}$. Since the set $\mathcal{D}$ is discrete, the traditional concept of continuous function is useless to assess this relationship of closeness. But observe that,

a) In models 1 and 2, we have $m_Y(0) = 1$, $m_Y(1) = 3$, $m_Y(2) = 5$, $m_Y(3) = 7$, $m_Y(4) = 9$ and so on. In these models, close values in $\mathcal{D}$ have fairly close conditional expectations and, as a result, if the sample size is small (as in Table 1), smoothing may improve the behaviour of the estimates. For a fixed sample size, the higher the variance of Z the better it will be to smooth: in this case, it will be likely to have points $\wp$ for which $Z = \wp$ in only a few observations and, then, smoothing will improve the accuracy of the nonparametric estimate. This is what we see when comparing models 1 and 2 in table 1: in the latter, we achieve by smoothing a comparatively more important improvement when we smooth.

b) In models 3, 4, 5 and 6, we have $m_Y(0) = -2$, $m_Y(1) = 2$, $m_Y(2) = 0$, $m_Y(3) = -8$, $m_Y(4) = -22$ and so on. Thus, close values in $\mathcal{D}$ do not have close conditional expectations. As a result, in no case is smoothing advisable. Even more, the higher the variance of Z, the worse it will be to smooth: if Z has small variance we will have plenty of information for each observed data point and the smoothing will not worsen dramatically the performance of the nonparametric estimate; however, if Z has large variance, then «noisy» information which comes from smoothing will seriously affect the performance of the nonparametric estimate. In tables 1, 2 and 3 we observe that in models 3-8 the non-smoothing estimate is the best one and the other nonparametric estimates only seem adequate in those models in which $Var(Z) = 0.3$.

To sum up, if the unknown part of the partially linear regression model does not exhibit high volatility, then the k-NN and the kernel estimates may perform slightly better than the non-smoothing one if the smoothing values are properly chosen. Otherwise, smoothing techniques are not adequate and may produce extremely misleading results, as in models 4, 6 and 8 –and observe that this may happen even though there exist continuous functions from $\mathbb{R}^q$ to $\mathbb{R}$ $m_Y(.)$ and $m_X(.)$ such that $\forall\, \wp \in \mathcal{D}$ $E[Y|Z = \wp] = m_Y(\wp)$ and $E[X|Z = \wp] = m_X(\wp)$. However, we must bear in mind that these results have been achieved with a fixed smoothing value. It is possible that data-dependent selections of h_n (or k_n) may alter these conclusions because, at the best of our knowledge, no theoretical result about it has already been developed in this context.

We have also generated observations from five pairs of regression curves with similar shape and computed the semiparametric estimates described in Section 3.2. In all cases Z and Z^* were taken as independent random variables from a Poisson distribution with mean λ (specified below), V and V^* were taken as independent random variables (also independent from Z and Z^*) from a normal distribution with zero mean and variance 1 and, finally, $\zeta = m(Z) + V$ and $\zeta^* = m^*(Z^*) + V^*$, where $m(Z)$ is specified below and $m^*(.)$ and $m(.)$ satisfy [23] for $\theta_0 = (10,2)$. The complete description of all models is as follows:

Model	9	10	11	12	13
λ	1.0	3.0	2.0	0.5	5.0
$m(Z)$	$2+Z$	$2+Z$	$(2-Z)^2$	$3\,(Z-1)^2$	$\log\,(Z+2)$

A trimming value ε had to be chosen in order to compute $\hat{\theta}$ and, additionally, positive real values ρ and α had to be selected to compute $\tilde{\theta}$. According to Theorem 6, the performance of $\hat{\theta}$ depends crucially on the choice of ε; according to Theorem 7, the performance of $\tilde{\theta}$ does not depend on the choice of ε, δ and α. In order to analyse how to choose ε, we have first computed in models 9-13 what values in F should satisfy $w(\wp) = 1$ to achieve as good an estimate of θ as possible. We obtained that these values

are: {0,1,2,3} in model 9, {1,2,3,4,5,6} in model 10, {0,1,2,3,4,5,6} in model 11, {0,1} in model 12 and {1,2,3,4,5,6,7,8,9} in model 13. Thus, we observe that the higher the variance of Z, the greater the number of values in F which must satisfy $w(\not{z}) = 1$ –and, hence, the smallest the positive real number ε should be. Therefore, in our simulations we have selected two values of ε which are inversely proportional to the standard deviation of Z. Specifically, we chose $\varepsilon_1 = 0.05 \times Var(Z)^{1/2}$ and $\varepsilon_2 = 0.1 \times Var(Z)^{1/2}$. With this choice, according to Theorems 5 and 6, the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta)$ is $N(0, \Sigma_1)$ for ε_1 and $N(0, \Sigma_2)$ for ε_2, and the asymptotic distribution of $n^{1/2}(\tilde{\theta} - \theta)$ is $N(0, \Sigma_3)$, where the symmetric matrices Σ_1, Σ_2 and Σ_3 for each model are as follows.

Model	9	10	11	12	13
Σ_1	$\begin{pmatrix} 80.3 & -26.0 \\ & 8.97 \end{pmatrix}$	$\begin{pmatrix} 75.0 & -13.7 \\ & 2.78 \end{pmatrix}$	$\begin{pmatrix} 11.5 & -3.25 \\ & 1.72 \end{pmatrix}$	$\begin{pmatrix} 16.5 & -5.50 \\ & 3.89 \end{pmatrix}$	$\begin{pmatrix} 260 & -133.4 \\ & 70.6 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 84.2 & -28.3 \\ & 10.2 \end{pmatrix}$	$\begin{pmatrix} 94.5 & -18.4 \\ & 3.84 \end{pmatrix}$	$\begin{pmatrix} 10.5 & -3.54 \\ & 2.50 \end{pmatrix}$	$\begin{pmatrix} 16.5 & -5.50 \\ & 2.75 \end{pmatrix}$	$\begin{pmatrix} 320 & -166.5 \\ & 88.5 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 50 & -15 \\ & 5 \end{pmatrix}$	$\begin{pmatrix} 46.6 & -8.33 \\ & 1.67 \end{pmatrix}$	$\begin{pmatrix} 7 & -1 \\ & 0.5 \end{pmatrix}$	$\begin{pmatrix} 10.6 & -2.50 \\ & 1.11 \end{pmatrix}$	$\begin{pmatrix} 161 & -82.6 \\ & 43.7 \end{pmatrix}$

We report in Table 4 the mean (M) and variance (V) of $\tilde{\theta}$ and $\hat{\theta}$ computed using non-smoothing weights for the nonparametric estimates and $\alpha = 0.01$, $\rho = 0.1$. In Table 5 we report corresponding results when the nonparametric estimates are computed using kernel weights (with Epanechnikov kernel) and $h = 1.2$. All reported values are based on $n = 40$ observations and $r = 10000$ replications.

TABLE 4
Non-smoothing estimates ($n = 40$, $r = 10000$)

		$\hat{\theta}_1$		$\hat{\theta}_2$		$\tilde{\theta}_1$		$\tilde{\theta}_2$	
		ε_1	ε_2	ε_1	ε_2	ε_1	ε_2	ε_1	ε_2
M. 9	M	9.969	9.658	2.025	2.141	10.059	10.062	1.990	1.989
	V	3.942	60.167	0.533	8.221	2.314	2.315	0.272	0.272
M. 10	M	10.267	10.235	1.948	1.953	10.266	10.270	1.948	1.947
	V	2.191	3.063	0.087	0.127	2.529	2.520	0.101	0.101
M. 11	M	10.090	10.067	1.968	1.976	10.073	10.073	1.969	1.968
	V	0.281	0.338	0.076	0.192	0.297	0.298	0.085	0.085
M. 12	M	9.985	9.801	2.010	1.999	10.032	10.032	1.993	1.992
	V	0.548	1.934	0.090	0.141	0.490	0.490	0.082	0.082
M. 13	M	12.491	12.530	0.690	0.660	12.554	12.560	0.650	0.647
	V	1.443	1.955	0.377	0.522	2.425	2.465	0.647	0.656

TABLE 5
Kernel estimates (n = 40, r = 10000)

		$\hat{\theta}_1$		$\hat{\theta}_2$		$\tilde{\theta}_1$		$\tilde{\theta}_2$	
		ε_1	ε_2	ε_1	ε_2	ε_1	ε_2	ε_1	ε_2
M. 9	*M*	9.746	9.583	2.093	2.155	9.804	9.807	2.071	2.070
	V	3.923	8.200	0.475	1.105	3.134	3.130	0.354	0.354
M. 10	*M*	10.020	9.971	1.998	2.008	10.013	10.015	1.999	1.999
	V	2.271	3.070	0.090	0.127	2.543	2.538	0.102	0.102
M. 11	*M*	10.089	10.053	1.974	1.989	10.067	10.068	1.982	1.981
	V	0.509	0.600	0.234	0.381	0.501	0.501	0.245	0.244
M. 12	*M*	10.008	9.719	1.974	2.024	10.059	10.062	1.955	1.953
	V	2.228	3.276	0.251	0.259	2.487	2.504	0.286	0.290
M. 13	*M*	11.670	11.708	1.125	1.100	11.732	11.733	1.087	1.087
	V	2.405	3.567	0.643	0.980	1.087	1.087	1.152	1.198

We observe that in models 9, 10, 11 and 12 the non-smoothing estimate performs better than the kernel one, whereas in model 13 the kernel estimate seems to be the most adequate one. Again, these results are not a surprise: in model 13 the regression function has low variability (i.e. close points in $\mathcal{D}$ have close conditional expectations) and as $Var(Z) = 5$ in every sample there are many different values –therefore, smoothing improves the accuracy of estimates.

If we compare $\hat{\theta}$ and $\tilde{\theta}$, we observe that, surprisingly, in some cases the former performs better than the latter (models 10 and 11 when ε_1 is used). This also happens with some other well-known two-stage estimators. The reason why this happens is because the weights $\lambda(\wp)$ are so poorly estimated in the first stage that no improvement is achieved in the second stage. However, in this specific model, the generalised least squares estimate still has an advantage over the ordinary least squares estimate: results do not depend on the choice of ε when using $\tilde{\theta}$, unlike what happens with $\hat{\theta}$ (see, for instance, models 9 and 12). This is the main reason why $\tilde{\theta}$ seems preferable to $\hat{\theta}$.

Appendix

Proof of Theorem 1: Given $\wp \in \mathcal{D}$, observe that $I(Z_j \neq \wp)\psi[Z_j - \wp)/h_n] = 0 \Rightarrow$

$$\tilde{W}_{nj}(\wp) = \{\psi[(\wp - Z_j)/h_n] \times [I(Z_j = \wp) + I(Z_j \neq \wp)]\}/\Sigma_k \psi[(\wp - Z_k)/h_n] = W_{nj}(\wp).$$

Therefore, $P\{\hat{m}_\zeta(\wp) \neq \tilde{m}_\zeta(\wp)\} \leq P\{\exists j: I(Z_j \neq \wp)\psi[(Z_j - \wp)/h_n] \neq 0\} = 1 -$

$$P\{I(Z \neq \wp) = 0 \text{ or } \psi[(Z - \wp)/h_n] = 0\}^n \leq 1 - P\{Z = \wp \text{ or } \|Z - \wp\| \geq h_n M\}^n = 1 - p_n(\wp)^n.$$

As [4] holds uniformly in $\mathcal{D}$, given $\varepsilon > 0$, $\exists\ n_0 : n \geq n_0 \Rightarrow$

$$n^t P\{\hat{m}_\zeta(Z) \neq \tilde{m}_\zeta(Z)\} \leq \sum_{\wp \in \mathcal{D}} nt\,(1 - p_n(\wp)^n) P(Z = z) \leq \sum_{\wp \in \mathcal{D}} \varepsilon P(Z = z) = \varepsilon. \blacksquare$$

Example 1: $P\{\hat{m}_\zeta(1) \neq \tilde{m}_\zeta(1)\} \geq$

$$P\{\textstyle\sum_j I(Z_j = 1) > 0,\ \sum_j \zeta_j I(Z_j = 1) \times \sum_j I(Z_j \neq 1)\psi[(Z_j - 1)/h_n] \neq$$

$$\textstyle\sum_j \zeta_j I(Z_j \neq 1)\psi[(Z_j - 1)/h_n] \times \sum_j I(Z_j = 1)\} =$$

$$= P\{\textstyle\sum_j I(Z_j = 1) > 0,\ \sum_j I(Z_j \neq 1)\psi[(Z_j - 1)/h_n] \neq 0\},$$

where the last equality holds by [5]. As $P(A \cap B) \geq 1 - P(A^c) - P(B^c)$, we have

$$P\{\hat{m}_\zeta(1) \neq \tilde{m}_\zeta(1)\} \geq 1 - P\{\textstyle\sum_j I(Z_j = 1) = 0\} -$$

$$P\{\textstyle\sum_j I(Z_j \neq 1)\psi[(Z_j - 1)/h_n] = 0\} = 1 - (1/2)^n - p_n\,(1)^n, \qquad \text{[A.1]}$$

where the last equality holds because $\psi(x) = 0 \Leftrightarrow ||x|| \geq 1$. Now, if we denote $S_n \equiv \{j \in \mathbb{N} : j > n^\gamma\}$, then

$$p_n\,(1)^n = (1 - c_0 \textstyle\sum_{j \in S_n} 1/j^2)^n \leq (1 - c_0 \sum_{j > n} 1/j^2)^n \equiv v_n.$$

The sequence v_n converges to $exp\{-c_0\}$ because

$$(1 - c_0 \textstyle\sum_{j>n} 1/j^2)^n = exp\{n \times log(1 - c_0 \sum_{j>n} 1/j^2)\} \sim exp\{n(-\ c_0 \sum_{j>n} 1/j^2)\},$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$. Now, $n(-\ c_0 \sum_{j>n} 1/j^2)$ converges to $-\ c_0$ as can be seen using Stolz criterion. Therefore $p_n\,(1)^n$ does not converge to 1. From [A.1] we deduce now that $P\{\hat{m}_\zeta(1) \neq \tilde{m}_\zeta(1)\}$ does not converge to 0 and, obviously, $P\{\hat{m}_\zeta(Z) \neq \tilde{m}_\zeta(Z)\}$ (which is greater than or equal to $(1/2) \times P\{\hat{m}_\zeta(1) \neq \tilde{m}_\zeta(1)\}$) does not converge to 0. ■

Proof of Theorem 2: We only prove the first statement here (the second one follows in a similar way). By [3] we know that $\exists\ M : ||x|| \geq M \Rightarrow \psi(x) = 0$, and there exists n_0 such that

$$n \geq n_0 \Rightarrow \mu/h_n \geq M \text{ and } ||\wp - Z_j||/h_n \geq MI(\wp \neq Z_j) \text{ if } \wp \in \mathcal{D},$$

where μ, defined in [6] is a fixed number. Hence if $n \geq n_0$ and $\wp \in \mathcal{D}$, then $\tilde{W}_{nj}(\wp) = W_{nj}(\wp)$ and $\tilde{m}_\zeta(\wp) = \hat{m}_\zeta(\wp)$. This result follows from this and from Theorem 1 in Delgado and Mora (1995). ■

Proof of Theorem 3: Given $\wp \in \mathcal{D}$, let us define $U_j(\wp) = (\zeta_j - m_\zeta(\wp))I(Z_j = \wp)$.

Then,

$$n^{1/2}\begin{pmatrix} \hat{m}_\zeta(\wp_1) - m_\zeta(\wp_1) \\ \cdots\cdots\cdots\cdots \\ \hat{m}_\zeta(\wp_f) - m_\zeta(\wp_f) \end{pmatrix} = (P_n \otimes I_s)^{-1}\, n^{-1/2} \sum_j \begin{pmatrix} U_j(\wp_1) \\ \cdots\cdots \\ U_j(\wp_f) \end{pmatrix},$$

where s = *dim* (ζ), I_s is the identity matrix of order s and P_n is the $f \times f$ diagonal matrix $P_n \equiv diag[n^{-1}\sum_j I(Z_j = \wp_1), \ldots, n^{-1}\sum_j I(Z_j = \wp_f)]$. Now, by Khinchine's Law of Large Numbers,

$$P_n \otimes I_s \xrightarrow{p} diag[p(\wp_1), \ldots, p(\wp_f)] \otimes I_s;$$

and by Linderberg-Levy's Central Limit Theorem

$$n^{-1/2}\sum_j \begin{pmatrix} U_j(\wp_1) \\ \cdots\cdots \\ U_j(\wp_f) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \begin{bmatrix} p(\wp_1)\Sigma(\wp_1) & \ldots & 0 \\ \cdots\cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & \ldots & p(\wp_f)\Sigma(\wp_f) \end{bmatrix}.$$

Combining both results we obtain Theorem 3. ■

Proof of Theorem 4: a) In the proof of Theorem 2 we show that, under these assumptions, there exists n_0 such that if $n \geq n_0$ and $\wp \in \mathcal{D}$, then $\tilde{m}_\zeta(\wp) = \hat{m}_\zeta(\wp)$. Thus, this result follows from this and from Theorem 2 in Delgado and Mora (1995).

b) The proof is similar to the proof of Theorem 2 in Delgado and Mora (1995), but now some changes have to be made.

$$n^{1/2}(\tilde{\beta} - \beta) = \tilde{\Phi}^{-1}(n^{-1/2}\textstyle\sum_i \tilde{\varepsilon}_{Xi}\, \tilde{\varepsilon}_{Ui} + n^{-1/2}\sum_i \tilde{\varepsilon}_{Xi}\, \tilde{\varepsilon}_{\theta i}). \qquad [A.2]$$

Note that this equation is similar to [A.2] in Delgado and Mora, but now there is a second term. The first term in the right-hand side of [A.2] behaves in the same way as the rightmost term in [A.2] in Delgado and Mora (1995): the same proof applies, except that now the universal consistency result we must refer to is Theorem 1 in Devroye and Wagner (1980) (universal consistency of kernel weights), instead of Theorem 1 in Delgado and Mora (1995) (universal consistency of non-smoothing weights with discrete regressors). As for the second term, it will suffice to prove that $E\|n^{-1/2}\sum_i \tilde{\varepsilon}_{Xi}\, \tilde{\varepsilon}_{\theta i}\| = o(1)$, but:

$$E\|n^{-1/2}\textstyle\sum_i \tilde{\varepsilon}_{Xi}\, \tilde{\varepsilon}_{\theta i}\| \leq n^{1/2} E[\|\tilde{\varepsilon}_{X1}\, \tilde{\varepsilon}_{\theta 1}\|] \leq$$

$$\leq E[\|X_1 - m_{X1}\|^2]^{1/2} E[n\|\tilde{\varepsilon}_{\theta 1}\|^2]^{1/2} + E[\|m_{X1} - \tilde{m}_{X1}\|^2]^{1/2} E[n\|\tilde{\varepsilon}_{\theta 1}\|^2]^{1/2}.$$

Thus, it only remains to prove that $E[n\|\tilde{\varepsilon}_{\theta 1}\|^2] = o(1)$. Now, $\tilde{\varepsilon}_{\theta 1} = \theta(Z_1) - \tilde{m}_{\theta(Z1)}$. Let A be the event $\{\tilde{m}_{\theta(Z1)} = \hat{m}_{\theta(Z1)},\ I_1 = 1\}$. If A is true, then, $\tilde{\varepsilon}_{\theta 1} = 0$. Therefore, $E[n\|\tilde{\varepsilon}_{\theta 1}\|^2] = nP(A^c) \times E[\|\theta(Z_1) - \tilde{m}_{\theta(Z1)}\|^2 \mid A^c]$; now, $nP(A^c)$ converges to 0 by Theorem 1 and the second term is bounded because $E[\theta(Z_1)^2] < \infty$. ■

Proof of Corollary 2: Under this assumptions, for n large enough $\tilde{m}_\zeta(\varphi) = \hat{m}_\zeta(\varphi)$; thus, we may use the non-smoothing estimate. As in Theorem 2 in Delgado and Mora (1995), it suffices to prove that

$$n^{-1/2} \sum_i \hat{\varepsilon}_{Xi} \hat{\varepsilon}_{Ui} I_i \xrightarrow{d} \mathcal{N}(0, \Psi), \qquad [A.3]$$

$$\hat{\Psi} \xrightarrow{p} \Psi, \qquad [A.4]$$

where $\Psi \equiv E\{\sigma^2(X, Z)(X - E[X|Z])(X - E[X|Z])'\}$. Both [A.3] and [A.4] follow in a similar way to [A.3] and [A.4] in Delgado and Mora (1995). For example, Proposition 2.1 still holds because equation [A.5] in Delgado and Mora (1995) is also true; the first term in this equation converges to 0 and the second one is 0 because

$$E[I_1 \hat{m}_{U1}(m_{X1} - \hat{m}_{X1})'(m_{X2} - \hat{m}_{X2}) \hat{m}_{U2} I_2] =$$

$$(n-2) \sum_{j,j\neq 1} \sum_{i,i\neq 2} E[I_1 \sigma^2(X_3, Z_3)(m_{X1} - X_j)'(m_{X2} - X_i) \hat{W}_{ji}(Z_1, Z_2, \ldots, Z_n) I_2],$$

where $\hat{W}_{ji}(Z_1, Z_2, \ldots, Z_n) \equiv W_{nj}(Z_1)\, W_{ni}(Z_2)\, W_{n3}(Z_1)\, W_{n3}(Z_2)$; all terms in this double summation are 0 because

$$E\{I_1 \hat{W}_{ji}(Z_1, Z_2, \ldots, Z_n) I_2 E[\sigma^2(X_3, Z_3)(m_{X1} - X_j)'(m_{X2} - X_i) | Z_1, \ldots, Z_n]\} =$$

$$E[I_1 \hat{W}_{ji}(Z_1, Z_2, \ldots, Z_n) I_2 \sigma^2(m_{X3}, Z_3)(m_{X1} - m_{Xj})'(m_{X2} - m_{Xi})] = 0.$$

Oddly enough, the moment condition required in both the homoskedastic model and the heteroskedastic one is the same. In the homoskedastic model, second order moments are required to prove [A.3] and fourth order moments are required to prove [A.4]; in the heteroskedastic model fourth order moments are required to prove both [A.3] and [A.4]. ■

Proof of Theorem 5: The following lemmas will be used in the proof. They are versions of Robinson's (1988) lemmas adapted to the mixed case. Throughout this proof, Robinson will mean Robinson (1988).

In the following lemmas, Z is a random variable which satisfies [9], $Z^{(2)}(d)$ denotes the conditional random variable $Z^{(2)}/Z^{(1)} = d$, f_d is the conditional probability density function of $Z^{(2)}(d)$, k is a function from $\mathbb{R}$ to $\mathbb{R}$ such that $\int |uk(u)|\, du < \infty$, ψ is a function from $\mathbb{R}^s$ to $\mathbb{R}$ defined by $\psi(u_1, \ldots, u_s) = k(u_1) \ldots k(u_s)$ and h_n is a sequence of positive real numbers. All notation here refers to that introduced in Section 3.1. after Corollary 2.

Lemma 1. If there exist real numbers M, M' such that $f_d(u) < M$ $(\forall u, \forall d \in \mathcal{Z})$ and $|k(u)| < M'$ $(\forall u \in \mathbb{R})$ then,

$$h(d, u) \equiv E[|\psi((Z^{(2)} - u)/h_n)| I(Z^{(1)} = d)] = O(h_n^s).$$

Proof: $h(d, u) = P(Z^{(1)} = d) \times E[|\psi((Z^{(2)} - u)/h_n)| \,|\, Z^{(1)} = d] =$

$$= P(Z^{(1)} = d) \times \int |\psi((v-c)/h_n)| f_d(v)dv \leq P(Z^{(1)} = d) \times M\left(\int |k(u)|\,du\right)^s h_n^s \leq Ch_n^s,$$

$$\text{where } C = M\left(\int |k(u)|\,du\right)^s < \infty. \blacksquare$$

Lemma 2. If there exist real numbers M, M' such that $f_d(u) < M$ $(\forall u,\ \forall d \in \mathcal{D})$ and $|k(u)| < M'$ $(\forall u \in \mathbb{R})$ and $g(d,u)$ is a function from $\mathbb{R}^q$ to $\mathbb{R}$ such that $E[|g(Z)|] < \infty$, then,

$$E[|g(Z_1)\psi_{12}(h_n)| I(Z_2^{(1)} = Z_1^{(1)})] = O(h_n^s).$$

Proof: If $h(.,.)$ is as defined in Lemma 1, then

$$E[|g(Z_1)K_{12}(a_n)| I(Z_2^{(1)} = Z_1^{(1)})] = E\{|g(Z_1)|E[|K_{12}(a_n)| I(Z_2^{(1)} = Z_1^{(1)})|Z_1]\} =$$

$$= E[|g(Z_1)|h(Z)] \leq Ca_n^q E[|g(Z_1)|] = C'a_n^q,$$

where $C' \equiv CE[|g(Z_1^{(1)}\ Z_1^{(2)})|] < \infty$ (the last inequality holds by Lemma 1). ■

Lemma 3. If $f_d \in G_\lambda^\infty$ and $k \in K_l$ $(l-1 < \lambda \leq l)$ then,

$$E\{(h_n^{-s} E[\psi_{12}(h_n)|Z_1^{(2)}] - f_d(Z_1^{(2)}))^2 | Z_1^{(1)} = d\} = O(h_n^{2\lambda})$$

Proof: Similar to Robinson's Lemma 4. ■

Lemma 4. Let $g(d,u)$ be as in Lemma 2 and define $g_d(u) = g(d,u)$. If there exist positive real numbers λ, α, μ such that $\forall d \in \mathcal{D}$ (and uniformly in d) $f_d \in G_\lambda^\infty(Z^{(2)}|Z^{(1)} = d)$, $g_d \in G_\mu^\alpha(Z^{(2)}|Z^{(1)} = d)$ and $k \in K_{l+m-1}$ (where $l-1 < \lambda \leq l$, $m-1 < \mu \leq m$ and $\eta = min(\mu, \lambda+1)$), then

$$E\{|E[(g(Z_1) - g(Z_2)\psi_{12}(a_n) I(Z_1^{(1)} = Z_2^{(1)})]\big|Z_1|^\alpha\} = O(h_n^{\alpha(s+\eta)})$$

Proof: Similar to proof of Lemma 2 applying Robinson's Lemma 5. ■

We can now prove Theorem 5. It will suffice to prove that

$$n^{-1/2} \sum_i (X_i - m^*_{Xi})(U_i - m^*_{Ui}) I_i \xrightarrow{d} N(0, \sigma^2 \Phi), \qquad [A.5]$$

$$n^{-1} \sum_i (X_i - m^*_{Xi})(X_i - m^*_{Xi})' I_i \xrightarrow{p} \Phi, \qquad [A.6]$$

$$n^{-1/2} \sum_i (X_i - m^*_{Xi})(\theta_i - m^*_{\theta i}) I_i \xrightarrow{p} 0, \qquad [A.7]$$

$$\sigma^{*2} \xrightarrow{p} \sigma^2. \qquad [A.8]$$

All of these results can be proved in a similar way to Robinson's propositions 1-15 though under our assumptions some of his propositions may be omitted and Cauchy-Schwarz inequality may be used instead. Lemmas in Robinson's appendix B do not apply any more; instead, the lemmas specified above must be used. ■

Proof of Theorem 6: By solving the optimization problem, we obtain

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \left\{ \sum_{\wp \in F} \hat{w}_n(\wp) \begin{pmatrix} 1 & \hat{m}(\wp) \\ \hat{m}(Z) & \hat{m}(\wp)^2 \end{pmatrix} \right\}^{-1} \sum_{\wp \in F} \hat{w}_n(\wp) \begin{pmatrix} \hat{m}^*(\wp) \\ \hat{m}(\wp)\,\hat{m}^*(\wp) \end{pmatrix}$$

Let us now consider the unfeasible estimate

$$\begin{pmatrix} \bar{\theta}_1 \\ \bar{\theta}_2 \end{pmatrix} = \left\{ \sum_{\wp \in F} W(Z) \begin{pmatrix} 1 & \hat{m}(\wp) \\ \hat{m}(\wp) & \hat{m}(\wp)^2 \end{pmatrix} \right\}^{-1} \sum_{\wp \in F} W(\wp) \begin{pmatrix} \hat{m}^*(\wp) \\ \hat{m}(\wp)\,\hat{m}^*(\wp) \end{pmatrix}$$

(It is unfeasible because $w(\wp)$ is unknown). First we prove.

Lemma 5: $P(\hat{\theta} \neq \bar{\theta}) = o(1)$.

Proof: Let us define the following subsets of F:

$F_1 = \{\wp \in F : P(Z = \wp) < \varepsilon\}$,

$F_2 = \{\wp \in F : P(Z = \wp) > \varepsilon \text{ and } P(Z^* = \wp) < \varepsilon\}$,

$F_3 = \{\wp \in F : P(Z = \wp) > \varepsilon \text{ and } P(Z^* = \wp) > \varepsilon\}$;

so, $F = F_1 \cup F_2 \cup F_3$ and hence, $P(\hat{\theta} \neq \bar{\theta}) \leq \sum_{\wp \in F} P(w(\wp) \neq \hat{w}_n(\wp)) = S_1 + S_2 + S_3$, where $S_i \equiv \sum_{\wp \in F_i} P(w(\wp) \neq \hat{w}_n(\wp))$. We prove that S_3 converges to 0 (the proof for S_1 and S_2 is similar). Let us define

$$\Xi \equiv \{p \in \mathbb{R} : \exists \wp \in \mathbb{R} \text{ such that } P(Z = \wp) = p \text{ or } P(Z^* = \wp) = p\}.$$

This set is closed in $\mathbb{R}$ (note that $0 \in \Xi$); as $\varepsilon \notin \Xi \Rightarrow \exists\, \delta > 0$ such that $(\varepsilon - \delta, \varepsilon + \delta) \cap \Xi = \varnothing$. Then, if $\wp \in F_3$, applying Chebychev inequality we have,

$$P(w(\wp) \neq \hat{w}_n(\wp)) \leq P(n^{-1} \Sigma_j I(Z_j = \wp) < \varepsilon) + P(n^{-1} \Sigma_j I(Z_j^* = \wp) < \varepsilon)$$

$$\leq P(|n^{-1} \Sigma_j I(Z_j = \wp) - P(Z = \wp)| > \delta) + P(|n^{-1} \Sigma_j I(Z_j^* = \wp) - P(Z^* = \wp)| > \delta)$$

$$\leq n^{-1} \delta^{-1} P(Z = \wp)(1 - P(Z = \wp)) + n^{-1} \delta^{-1} P(Z^* = \wp)(1 - P(Z^* = \wp)).$$

Hence, $S_3 \leq 2/n\delta = o(1)$. ■

Now we prove Theorem 6: let us denote $\hat{v}(\wp) \equiv \hat{m}^*(\wp) - \theta_{10} - \theta_{20}\,\hat{m}(\wp)$. Then

$$n^{1/2} \begin{pmatrix} \bar{\theta}_1 - \theta_{10} \\ \bar{\theta}_2 - \theta_{20} \end{pmatrix} = \left\{ \sum_{\wp \in F} W(\wp) \begin{pmatrix} 1 & \hat{m}(\wp) \\ \hat{m}(\wp) & \hat{m}(\wp)^2 \end{pmatrix} \right\}^{-1} \times n^{1/2} \sum_{\wp \in F} W(\wp) \begin{pmatrix} 1 \\ \hat{m}(\wp) \end{pmatrix} \hat{v}(\wp),$$

where now both summations run only through a finite number of terms which does not depend on n. Now, by Theorem 3

$$\sum_{\wp \in F} {}^{W(\wp)} \begin{pmatrix} 1 & \hat{m}(\wp) \\ \hat{m}(\wp) & \hat{m}(\wp)^2 \end{pmatrix} - A = o_p(1).$$

On the other hand, if $f = \#\{\wp \in F : w(\wp) = 1\}$ and we denote $\wp_1, \ldots, \wp_f$ the points in F satisfying that $w(\wp) = 1$, then

$$n^{1/2} \sum_{\wp \in F} {}^{W(\wp)} \begin{pmatrix} 1 \\ \hat{m}(\wp) \end{pmatrix} \hat{v}(\wp) = \begin{pmatrix} 1 & \ldots & 1 \\ \hat{m}(\wp_1) & \ldots & \hat{m}(\wp_f) \end{pmatrix} \times n^{1/2} \begin{pmatrix} \hat{v}(\wp_1) \\ \vdots \\ \hat{v}(\wp_f) \end{pmatrix}.$$

Now, as $\hat{v}(\wp) = (\hat{m}^*(\wp) - m^*(\wp)) + \theta_{20}(\hat{m}(\wp) - m(\wp))$, and the random samples in which each nonparametric estimate is based are independent, by Theorem 3

$$n^{1/2} \begin{pmatrix} \hat{v}(\wp_1) \\ \vdots \\ \hat{v}(\wp_f) \end{pmatrix} \xrightarrow{d} \mathcal{N} \left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \text{diag}\,[\lambda(\wp_1), \ldots, \lambda(\wp_f)] \right].$$

and hence Theorem 6 follows for the unfeasible estimate $\bar{\theta}$ and, applying Lemma 5, also for the feasible estimate $\hat{\theta}$. ■

Proof of Theorem 7: By solving the optimization problem we obtain

$$n^{1/2} \begin{pmatrix} \tilde{\theta}_1 - \theta_{10} \\ \tilde{\theta}_2 - \theta_{20} \end{pmatrix} = \left\{ \sum_{\wp \in F} \begin{pmatrix} 1 & \hat{m}(\wp) \\ \hat{m}(\wp) & \hat{m}(\wp)^2 \end{pmatrix} \hat{\lambda}(\wp)^{-1} \hat{u}_n(\wp) \right\}^{-1}$$

$$\times n^{1/2} \sum_{\wp \in F} \begin{pmatrix} 1 \\ \hat{m}(\wp) \end{pmatrix} \hat{v}(\wp) \hat{\lambda}(\wp)^{-1} \hat{u}_n(\wp)$$

where $\hat{v}(\wp)$ is as in Theorem 6. Thus it suffices to prove that

$$\sum_{\wp \in F} (1, \hat{m}(\wp))'(1, \hat{m}(\wp)) \hat{\lambda}(\wp)^{-1} \hat{u}_n(\wp) \xrightarrow{p} \Omega, \qquad [A.9]$$

$$n^{1/2} \sum_{\wp \in F} (1, \hat{m}(\wp))' \hat{v}(\wp) \hat{\lambda}(\wp)^{-1} \hat{u}_n(\wp) \xrightarrow{d} \mathcal{N}(0, \Omega). \qquad [A.10]$$

We prove [A.10]; [A.9] follows in a similar way.

$n^{1/2} \sum_{\wp \in F} (1, \hat{m}(\wp))' \hat{v}(\wp) \hat{\lambda}(\wp)^{-1} \hat{u}_n(\wp) = T_1 + T_2 + T_3 + T_4 + T_5$, where

$T_1 = n^{1/2} \sum_{\wp \in F} (1, m(\wp))' \hat{v}(\wp) (\hat{\lambda}(\wp)^{-1} - \lambda(\wp)^{-1} \hat{u}_n(\wp)$

$T_2 = n^{1/2} \sum_{\wp \in F} (0, \hat{m}(\wp) - m_\zeta(\wp))' \hat{v}(\wp) \lambda(\wp)^{-1} \hat{u}_n(\wp)$

$$T_3 = n^{1/2}\Sigma_{\wp\in F}\ (1, m_\zeta(\wp))'\hat{v}(\wp)\ \lambda(\wp)^{-1}\ (\hat{u}_n(\wp) - u_n(\wp))$$

$$T_4 = n^{1/2}\Sigma_{\wp\in F}\ (1, m_\zeta(\wp))'\ (1, \theta_{20})\{\hat{\Pi}(\wp)^{-1} - \Pi(\wp)^{-1}\}\varphi(\wp)\lambda(\wp)^{-1}u_n(\wp)$$

$$T_5 = n^{1/2}\Sigma_{\wp\in F}\ (1, m_\zeta(\wp))'\ (1, \theta_{20})\,\Pi(\wp)^{-1}\varphi(z)\lambda(\wp)^{-1}u_n(\wp)$$

where we denote $u_n(\wp) = I(P(Z^* = \wp) \geq \rho/n^\alpha) \times I(P(Z = \wp) \geq \rho/n^\alpha)$,

$$\hat{\Pi}(\wp) = \begin{pmatrix} \Sigma_j I(Z_j^* = \wp)/n & 0 \\ 0 & \Sigma_j I(Z_j = \wp)/n \end{pmatrix}, \qquad \Pi(\wp) = \begin{pmatrix} P(Z^* = \wp) & 0 \\ 0 & P(Z = \wp) \end{pmatrix},$$

$$\varphi(\wp) = \begin{pmatrix} \Sigma_j(\zeta_j^* - m^*(\wp))I(Z^*_j = \wp)/n \\ \Sigma_j(\zeta_j - m(\wp))I(Z_j = \wp)/n \end{pmatrix}.$$

Now, $T_5 \xrightarrow{d} \mathcal{N}(0, \Omega)$ because if we define

$$X_{nj} = n^{1/2}\Sigma_{\wp\in F}\ (1, m_\zeta(\wp))'\ (1, \theta_{20})\,\Pi(\wp)^{-1}\begin{pmatrix} (\zeta_j^* - m_\zeta^*(\wp))I(Z^*_j = \wp)/n \\ (\zeta_j - m_\zeta(\wp))I(Z_j = \wp)/n \end{pmatrix}\lambda(\wp)^{-1}u_n(\wp)$$

then $T_5 = \Sigma_j X_{nj}$ and X_{nj} is a triangular array with independent random variables within rows which satisfies the Lindeberg condition (see e.g. Serfling 1980, Section 1.9.3.). Using similar arguments, it is easily proved that $T_i \xrightarrow{p} 0$, for $1 \leq i \leq 4$. ■

References

Bierens, H. J. (1987): «Kernel estimators of regression functions», *Advances in Econometrics, Fifth World Congress,* Vol. I, T. F. Bewley (ed.), Cambridge: Cambridge University Press, pp. 99-144.

Carroll, R. J. (1982): «Adapting for heteroscedasticity in linear models», *Annals of Statistics* 10, pp. 1.224-1.233.

Chamberlain, G. (1986): «Asymptotic efficiency in semiparametric models with censoring», *Journal of Econometrics* 32, pp. 189-218.

Chamberlain, G. (1992): «Efficiency bounds for semiparametric regression», *Econométrica* 60, pp. 567-596.

Chen, H. (1988): «Convergence rates for parametric components in a partly linear model», *Annals of Statistics* 16, pp. 136-146.

Chen, H. and Shiau, J. H. (1991): «A two-stage spline smoothing method for partially linear models», *Journal of Statistics, Planning and Inference* 27, pp. 187-201.

Delgado, M. A. (1992): «Semiparametric generalised least squares in the multivariate nonlinear regression model», *Econometric Theory* 8, pp. 203-222.

Delgado, M. A. and Mora, J. (1995): «Nonparametric and semiparametric estimation with discrete regressors», *Econometrica* 63, 1477-1484.

Delgado, M. A. and Stengos, T. (1994): «Semiparametric specification testing of non-nested econometric models», *Review of Economic Studies* 207 pp. 291-303.

Devroye, L. (1978): «The uniform convergence of nearest neighbor regression function estimators and their application in optimization», *IEEE Transactions on Information Theory* IT-24, pp. 142-151.

Devroye, L. and Wagner, T. J. (1980): «Distribution-free consistency results in nonparametric discrimination and regression function estimation», *Annals of Statistics* 8, pp. 231-239.

Eicker, F. (1963): «Asymptotic normality and consistency of the least squares estimator for families of linear regressions», *Annals of Mathematical Statistics* 34, pp. 447-456.

Engle, R. F.; Granger, W. J.; Rice, J. A. and Weiss, A. (1986): «Semiparametric estimates of the relationship between weather and electricity sales», *Journal of the American Statistical Association* 81, pp. 310-320.

Gasser, T.; Müller, H. G.; Köhler, W.; Molinari, L. and Prader, A. (1984): «Nonparametric regression analysis of growth curves», *Annals of Statistics* 12, pp. 210-229.

Härdle, W. and Marron, J. S. (1990): «Semiparametric comparison of regression curves», *Annals of Statistics* 18, pp. 63-89.

Härdle, W. and Stoker, T. M. (1989): «Investigating smooth multiple regression by the method of average derivatives», *Journal of the American Statistical Association* 84, pp. 986-995.

Heckman, N. E. (1986): «Spline smoothing in a partly linear model», *Journal of the Royal Statistical Society* B, 48, pp. 244-248.

Lawton, W. H.; Sylvestre, E. A. and Maggio, M. S. (1972): «Self modeling nonlinear regression», *Technometrics* 14, pp. 513-532.

Nadaraya, E. A. (1964): «On estimating regression», *Theory of Probability and its Applications* 9, pp. 141-142.

Newey, W. K. (1990): «Efficient instrumental variable estimation of nonlinear models», *Econometrica* 58, pp. 809-837.

Pinkse C. A. P. and Robinson, P. M. (1995): «Pooling nonparametric estimates of regression functions with a similar shape», in *Advances in Econometrics and Quantitative Economics*, G. S. Maddala et al. (eds.), Basil Blackwell, pp. 172-197.

Powell, J. L.; Stock, J. L. and Stoker, T. M. (1989): «Semiparametric estimation of index coefficients», *Econometrica* 57, pp. 1.403-1.430.

Rice, J. A. (1986): «Convergence rates for partially splined models», *Statistics and Probability Letters* 4, pp. 203-208.

Robinson, P. M. (1987): «Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form», *Econometrica* 55, pp. 531-548.

Robinson, P. M. (1988), «Root-n-consistent semiparametric regression», *Econométrica* 56, pp. 931-954.

Robinson, P. M. (1989): «Hypothesis testing in nonparametric and semiparametric models for economic time series», *Review of Economic Studies* 56, pp. 511-534.

Robinson, P. M. (1993): «Nearest-neighbour estimation of semiparametric regression models», Manuscript, London School of Economics.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Speckman, P. (1988): «Kernel smoothing in partially linear models», *Journal of the Royal Statistical Society* B, 50, pp. 413-446.

Stocker, T. M. (1991): *Lectures on Semiparametric Econometrics* CORE Lecture Series, Universite Catholique de Louvain.

Stone, C. J. (1977): «Consistent nonparametric regression», *Annals of Statistics* 4, pp. 595-645.

Watson, G. S. (1964): «Smooth regression analysis, *Sankhya A* 26, pp. 359-372.
White, H. (1980): «A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity», *Econometrica* 48, pp. 817-838.

Resumen

Este artículo analiza la estimación de modelos no paramétricos y semiparamétricos cuando hay algún regresor discreto. Si todos los regresores son discretos, Delgado y Mora (1995) prueban que un sencillo estimador que no suaviza proporciona estimadores globalmente consistentes de la función de regresión. Aquí se sugiere que, en determinadas circunstancias, es aconsejable utilizar estimadores con suavizado. Se demuestra que los estimadores con suavizado más utilizados, como los estimadores núcleo o *k*-puntos más próximos, son asintóticamente equivalentes al estimador sin suavizado. También se considera el caso mixto en el que hay regresores discretos y continuos. Los estimadores no paramétricos que se presentan resultan útiles en muchos modelos semiparamétricos. Se describen en detalle el modelo de regresión parcialmente lineal y la modelización de curvas con forma similar. El artículo contiene también un estudio de simulación.

Recepción del original, junio de 1995
Versión final, septiembre de 1995