Local likelihood estimation of truncated regression and its partial derivatives: theory and application

Park, Byeong U.; Simar, Léopold; Zelenyuk, Valentin

Postprint / Postprint
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
www.peerproject.eu

Accepted Manuscript

Local likelihood estimation of truncated regression and its partial derivatives: Theory and application

Byeong U. Park, Léopold Simar, Valentin Zelenyuk

Please cite this article as: Park, B.U., Simar, L., Zelenyuk, V., Local likelihood estimation of truncated regression and its partial derivatives: Theory and application. *Journal of Econometrics* (2008), doi:10.1016/j.jeconom.2008.08.007

Local Likelihood Estimation of Truncated Regression and Its Partial Derivatives: Theory and Application

Byeong U. Park[*]
Department of Statistics
Seoul National University

Léopold Simar[†]
Institut de Statistique
Université Catholique de Louvain

Valentin Zelenyuk
Kyiv Economics Institute and
Kyiv School of Economics, Ukraine

Final version
August 05, 2008

Abstract

In this paper we propose a very flexible estimator in the context of truncated regression that does not require parametric assumptions. To do this, we adapt the theory of local maximum likelihood estimation. We provide the asymptotic results and illustrate the performance of our estimator on simulated and real data sets. Our estimator performs as good as the fully parametric estimator when the assumptions for the latter hold, but as expected, much better when they do not (provided that the curse of dimensionality problem is not the issue). Overall, our estimator exhibits a fair degree of robustness to various deviations from linearity in the regression equation and also to deviations from the specification of the error term. So the approach shall prove to be very useful in practical applications, where the parametric form of the regression or of the distribution is rarely known.

Key words : Nonparametric Truncated Regression, Local Likelihood.
JEL Classification: C14, C24

1   Introduction

In this paper we consider the problem of estimating a regression model where the support of the continuous dependent variable is bounded at a known constant at one end and when many of the observations are observed near this bound. This is a common case when the dependent variable is an economic index measured within some range. One example in the recent econometric literature of such context is the analysis of how some economic variables determine the level of the Debreu (1951)-Farrell (1957) type efficiency score (bounded between 0 and 1 or, taking its inverse, between 1 and infinity, with many values concentrated near unity). A similar example can be drawn for the applied consumer analysis, where the so-called Luenberger (1994)-Chambers *et al.* (1996) shortage/benefit or directional distance function (bounded between zero and infinity, with most values being close to zero) can be used to analyze consumer benefits. An appropriate way of handling such problems is to use the *truncated regression* approach (see Simar and Wilson, 2007, for a parametric case).

The traditional truncated regression approach is based on using fully specified parametric model, where both the functional form of the relationship between the dependent and explanatory variables and the functional form of the distribution of the error term is specified. A natural estimator therefore is based on the maximum likelihood principle. An obvious drawback of such approach is the questionable reliability of parametric assumptions and vulnerability to deviations from them. Indeed, a mistake in specifying a parametric form of the regression equation or of the distribution of the error may lead to inconsistent estimation. The goal of our study is to propose a more flexible estimator for the context of truncated regression that does not require such parametric assumptions.

In particular, we adapt the theory of local maximum likelihood estimation (*e.g.*, see Tibshirani and Hastie, 1987, Fan and Gijbels, 1996, Fan *et al.*, 1996, and Kumbhakar *et al.*, 2007) to the truncated regression case. Non-parametric approach to truncated regression was already investigated by Lewbel and Linton (2000), who used local least squares theory to address the problem. As defined, *e.g.* in Maddala (2001, p336), in truncated regression models, the dependent variable is either unobserved beyond some known threshold, or not defined beyond that threshold. In some cases, most of the observations tend towards this threshold, *e.g.*, as firms tend to be perfect and achieve the 100% efficiency level, but cannot be "over-perfect-efficient" by definition. In the regression context, ignoring the boundary or treating it as censoring threshold may result in serious bias and inconsistency of the estimates of the fitted values and of the marginal effects (see Simar and Wilson, 2007 for extensive discussions on this in a parametric setup). On the other hand, mis-specifying the regression equation or/and the error (*e.g.*, wrong guess about heteroskedasticity) may also lead to

1

biased and inconsistent estimation. The flexibility offered by the local likelihood methods help circumventing these problems substantially, as we demonstrate with some simulated examples.

The theoretical foundation for our paper is based on a recent paper of Kumbhakar *et al.* (2007) who extended and generalized the approach suggested by Fan *et al.* (1996). In our work we make further extension. First, we adapt the theory to the truncated regression case. Second, and most importantly, we provide asymptotic results for the derivatives of regression function, which is the main focus of our paper, because many economic studies are concerned with the marginal effects of some variables on others. Third, our treatment includes both the cases where the shape parameter of the error distribution is an unknown constant, and where it is an unknown smooth function. In the former case, our estimator of the shape parameter achieves root-n consistency, and so does not suffer from the curse of dimensionality. Fourth, we show that fitting a lower order polynomial for the shape parameter may jeopardize the estimator of the regression function. This justifies consideration of higher-order local polynomial fit for the shape parameter even if one is mainly interested in estimating the regression function and its derivatives.

Our paper is structured as follows. In Section 2, we describe the local likelihood truncated regression methods. In Section 3, we present the asymptotic theory of local likelihood adapted to the truncated regression case of our type. In Section 4, we illustrate the performance of our estimator on several simulated data sets, considering different scenarios about regression equation and the error. In Section 5, we illustrate our estimator for a real data set. Section 6 concludes and Section 7 gives the regularity conditions for obtaining our results and outlines the proof of the theorems.

2 Local Polynomial Estimation

2.1 Constant Shape Parameter Case

We observe a set of i.i.d. random variables (X_i, Y_i) for $i = 1, \ldots, n$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$, where

$$Y_i = f(X_i) + \varepsilon_i \geq c$$

for some unknown function f and a known positive constant c. In this model, ε, conditionally on $X = x$, has a continuous distribution $G(\cdot, \tau)$ truncated below $c - f(x)$, where τ is an unknown shape parameter that is assumed to be a constant. In other words, the conditional density of Y given $X = x$ equals

$$\varphi(y, f(x), \tau) \equiv \frac{g_\varepsilon\left(y - f(x), \tau\right)}{1 - G\left(c - f(x), \tau\right)}\, I(y \geq c),$$

where $g_\varepsilon(\varepsilon, \tau) = \partial G(\varepsilon, \tau)/\partial \varepsilon$. We assume that G is known. For instance, if G is the normal distribution, and if we impose some parametric (*e.g.* linear) model for the regression function, we have the classical truncated regression models as in Maddala (2001, p337). Our main interest here is estimation of the function f and its derivatives. Below we describe local polynomial estimation of f in a general setting of multivariate X.

Note that all the results obtained in this paper could be easily adapted to the case where the dependent variable is truncated at both sides : $c_1 \le Y_i \le c_2$. This would only change the definition of $\varphi(y, f(x), \tau)$, the conditional density of Y given $X = x$.

Define $\ell = \log \varphi$. Then, the conditional log-likelihood equals $\sum_{i=1}^n \ell(Y_i, f(X_i), \tau)$. Let x be a point at which one wants to estimate the values of the function f and its derivatives. A local conditional log-likelihood is obtained by replacing f in the conditional log-likelihood by its pth order polynomial approximation in a neighborhood of x and putting the weight $K_h(X_i - x)$ for each observation (X_i, Y_i), where $K_h(u) = h^{-d}K(h^{-1}u)$, K is a d-variate kernel function, typically a symmetric density function defined on $\mathbb{R}^d$, and h is a positive scalar, called the bandwidth. Precisely, it is given by

$$
\begin{aligned}
&L_n(\theta_0, \theta_1, \ldots, \theta_{r(p)-1}, \tau; x) \\
&= \sum_{i=1}^n \ell\left(Y_i, \theta_0 + \theta_1(X_{i1} - x_1) + \cdots + \theta_{r(p)-1}(X_{id} - x_d)^p, \tau\right) K_h(X_i - x),
\end{aligned}
$$

where $r(p)-1$ is the total number of partial derivatives up to order p, i.e., $r(p) = \sum_{j=0}^p \binom{j+d-1}{d-1}$. Here and below, $X_i \equiv (X_{i1}, \ldots, X_{id})^T$ and $x \equiv (x_1, \ldots, x_d)^T$. The pth order local polynomial estimators of f and its derivatives at x are obtained by maximizing $L_n(\theta_0, \theta_1, \ldots, \theta_{r(p)-1}, \tau; x)$. For example, $\hat{f}(x) = \hat{\theta}_0(x)$ and the estimator of $f'(x) \equiv [\partial f(x)/\partial x_1, \ldots, \partial f(x)/\partial x_d]^T$ is given by $\hat{f}'(x) = [\hat{\theta}_1(x), \ldots, \hat{\theta}_d(x)]^T$, where

$$
\left(\hat{\theta}_0(x), \hat{\theta}_1(x), \ldots, \hat{\theta}_{r(p)-1}(x), \tilde{\tau}(x)\right) = \arg\max_{\theta_0, \ldots, \theta_{r(p)-1}, \tau} L_n(\theta_0, \theta_1, \ldots, \theta_{r(p)-1}, \tau; x). \quad (2.1)
$$

The estimator $\tilde{\tau}$ is obtained locally in the above local polynomial estimation procedure, and thus it depends on x. It can be improved by maximizing the full likelihood with f being replaced by its estimator $\hat{f}$, i.e, a better estimator is given by

$$
\hat{\tau} = \arg\max_\tau \sum_{i=1}^n \ell(Y_i, \hat{f}(X_i), \tau). \quad (2.2)
$$

One may further update the estimators $\hat{\theta}_j(x)$ by maximizing $L_n(\theta_0, \theta_1, \ldots, \theta_{r(p)-1}, \hat{\tau}; x)$ where τ on the right hand side of (2.1) is replaced by $\hat{\tau}$, now with respect to $\theta_0, \theta_1, \ldots, \theta_{r(p)-1}$ only.

3

2.2 Functional Shape Parameter Case

In this subsection we discuss the local likelihood truncated regression when the shape parameter τ is also a smooth function. A preliminary experiment showed that when the shape parameter is not a constant, fitting a local constant for τ as in the previous subsection produced poor estimates of f and its derivatives. This motivated us to consider fitting a higher order local polynomial for the function τ. As illustrated in the simulation study reported in Section 4 below, fitting a local linear for τ worked particularly well.

When the shape parameter is a function, the conditional log-likelihood is given by $\sum_{i=1}^{n} \ell(Y_i, f(X_i), \tau(X_i))$, where $\ell(y, \nu, \omega) = \log\left[g_\varepsilon(y - \nu, \omega)I(y \geq c)/\left\{1 - G(c - \nu, \omega)\right\}\right]$. We fit a qth order local polynomial for the function τ, i.e., we take the following local conditional log-likelihood:

$$
\begin{aligned}
L_n&(\theta_0, \ldots, \theta_{r(p)-1}, \tau_0, \ldots, \tau_{r(q)-1}; x) \\
&= \sum_{i=1}^{n} \ell\left(Y_i, \theta_0 + \theta_1(X_{i1} - x_1) + \cdots + \theta_{r(p)-1}(X_{id} - x_d)^p,\right. \\
&\qquad\qquad \left.\tau_0 + \tau_1(X_{i1} - x_1) + \cdots + \tau_{r(q)-1}(X_{id} - x_d)^q\right) K_h(X_i - x).
\end{aligned}
$$

The local polynomial estimators of f, τ and their derivatives at x are obtained by maximizing $L_n(\theta_0, \theta_1, \ldots, \theta_{r(p)-1}, \tau_0, \ldots, \tau_{r(q)-1}; x)$, i.e.,

$$
\begin{aligned}
\left(\hat{\theta}_0(x), \ldots, \hat{\theta}_{r(p)-1}(x), \hat{\tau}_0(x), \ldots, \hat{\tau}_{r(q)-1}(x)\right) & \qquad (2.3)\\
= \arg \max_{\theta_0, \ldots, \theta_{r(p)-1}, \tau_0, \ldots, \tau_{r(q)-1}} L_n(\theta_0, \theta_1, \ldots, \theta_{r(p)-1}, \tau_0, \ldots, \tau_{r(q)-1}; x). &
\end{aligned}
$$

3 Theoretical Properties

Here, we provide the asymptotic distributions of the estimators defined at (2.1)–(2.3). The theory we present here does not rely on the assumption that the log-likelihood function $\ell(y, \nu, \omega)$ as a function of (ν, ω) is *globally* concave for each y. The latter assumption is usually imposed for methods based on the local likelihood approach, see Fan *et al.* (1995) or Carroll *et al.* (1997), for example.

3.1 Constant Shape Parameter Case

The results we present below is closely related to those of Kumbhakar *et al.* (2007). However, the latter treated only the local linear estimator for multivariate X in some different setting. We give more general results for the local polynomial estimators defined at (2.1). Also, we show that the estimator of τ defined at (2.2) is $\sqrt{n}$-consistent.

For $0 \le i, j \le 2$ with $i + j = 1, 2$, let

$$\ell_{ij}(y, \nu, \omega) = \partial^{i+j} \ell(y, \nu, \omega) / (\partial \nu^i \partial \omega^j).$$

For a d-vector u, let $z_p(u) = (1, u_1, \ldots, u_d^p)^T$, an $r(p)$-vector. For a vector $a \equiv (a_0, \ldots, a_{r(p)-1})^T$ and a scalar b, define $Q(a, b) = [Q_1(a, b)^T, Q_2(a, b)]^T$ where

$$
\begin{aligned}
Q_1(a, b) &= \int E\left[\ell_{10}\left(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1} u_d^p, \tau + b\right) \Big| X = x\right] z_p(u) K(u) \, du, \\
Q_2(a, b) &= \int E\left[\ell_{01}\left(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1} u_d^p, \tau + b\right) \Big| X = x\right] K(u) \, du.
\end{aligned}
$$

Our results require that the system of equations $Q(a, b) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}$. As in Kumbhakar *et al.* (2007), the uniqueness of the solution plays an important role for a stochastic expansion of the estimators.

To give a sufficient condition for the uniqueness of the solution, define for $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$

$$\rho_{ij}(x, \alpha) = -E\left[\ell_{ij}(Y, f(x) + \alpha_1, \tau + \alpha_2) \,|\, X = x\right].$$

It can be shown that if

$$C_1(x) \equiv \inf_{\alpha \in \mathbb{R}^2} \rho_{20}(x, \alpha) > 0, \quad C_2(x) \equiv \inf_{\alpha \in \mathbb{R}^2} \left\{\rho_{02}(x, \alpha) - \frac{\rho_{11}(x, \alpha)^2}{\rho_{20}(x, \alpha)}\right\} > 0, \qquad (3.1)$$

and $K \ge 0$ is supported on a set which contains a d-dimensional open rectangle, then $Q(a, b) = 0$ has the unique solution. To see this, observe that a Taylor expansion gives

$$
\begin{aligned}
Q(a, b) &= -\int \begin{bmatrix} \rho_{20}(x, \alpha(u)) z_p(u) z_p(u)^T & \rho_{11}(x, \alpha(u)) z_p(u) \\ \rho_{11}(x, \alpha(u)) z_p(u)^T & \rho_{02}(x, \alpha(u)) \end{bmatrix} K(u) \, du \begin{pmatrix} a \\ b \end{pmatrix} \\
&\equiv -\mathcal{M} \begin{pmatrix} a \\ b \end{pmatrix},
\end{aligned}
$$

where $\alpha(u)$ lies on a line segment joining $(0, 0)$ and $(a_0 + a_1 u_1 + \cdots + a_{r(p)-1} u_d^p, b)$. Now, for any vector $c_1 \in \mathbb{R}^{r(p)}$ and any scalar $c_2 \in \mathbb{R}$, it follows that

$$
\begin{aligned}
(c_1^T, c_2) \mathcal{M} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} &= \int \rho_{20}(x, \alpha(u)) \left\{c_1^T z_p(u) + c_2 \frac{\rho_{11}(x, \alpha(u))}{\rho_{20}(x, \alpha(u))}\right\}^2 K(u) \, du \\
&\quad + \int c_2^2 \left\{\rho_{02}(x, \alpha(u)) - \frac{\rho_{11}(x, \alpha(u))^2}{\rho_{20}(x, \alpha(u))}\right\} K(u) \, du \\
&\equiv \mathcal{I}_1 + \mathcal{I}_2 \ge 0
\end{aligned}
$$

since $K \ge 0$. Thus, $\mathcal{M}$ is nonnegative definite. Suppose that

$$(c_1^T, c_2) \mathcal{M} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = 0.$$

Then, $\mathcal{I}_1 = \mathcal{I}_2 = 0$. Since

$$0 = \mathcal{I}_2 \geq C_2(x)\, c_2^2 \int K(u)\, du$$

we obtain $c_2 = 0$ from (3.1). Furthermore, the first integral with $c_2 = 0$ reduces to

$$\int \rho_{20}(x, \alpha(u)) \left\{ c_1^T z_p(u) \right\}^2 K(u)\, du \geq C_1(x) \int \left\{ c_1^T z_p(u) \right\}^2 K(u)\, du.$$

Thus, $\mathcal{I}_1 = 0$ implies that

$$\int \left\{ c_1^T z_p(u) \right\}^2 K(u)\, du = 0.$$

This means that $c_1^T z_p(u) = 0$ for all u in the support of K. Since a polynomial of order j has at most j roots, it implies that $c_1 = 0$. This shows that the matrix $\mathcal{M}$ is positive definite, from which we conclude that $Q(a, b) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}$.

Now, we present the asymptotic distributions of the estimators defined at (2.1) and (2.2). The presentation needs some careful notations to treat the multivariate X and the higher order approximating polynomial. First, for a d-tuple $k \equiv (k_1, \ldots, k_d)$ and a d-vector x, write

$$k! = k_1! \times \cdots \times k_d!, \quad |k| = \sum_{i=1}^{d} k_i, \quad x^k = x_1^{k_1} \times \cdots \times x_d^{k_d}.$$

For a function η defined on $\mathbb{R}^d$, write

$$\left(\nabla^k \eta \right)(x) = \frac{\partial^{|k|} \eta(x)}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}.$$

Let $m_j = \binom{j+d-1}{d-1}$ for $j \geq 0$. Arrange m_j number of d-tuples k with $|k| = j$ in a counter-lexicographical order: put $(j, 0, \ldots, 0)$ first and $(0, 0, \ldots, j)$ last. Let ξ_j denote the function which maps an integer s for $1 \leq s \leq m_j$ to the one located at the sth position in the arrangement of the d-tuples of size j. For example, $\xi_j(1) = (j, 0, \ldots, 0)$. Let $\mu_k = \int u^k K(u)\, du$ for a d-tuple k. For $j, l \geq 0$ denote by N_{jl} the $m_j \times m_l$ matrix whose (s, t)th entry equals $\mu_{\xi_j(s) + \xi_l(t)}$. Define $r(p) \times m_j$ matrices $N_j^{(p)} = \left(N_{0j}^T, \ldots, N_{pj}^T \right)^T$ for $j = 0, \ldots, p+1$, and a $r(p) \times r(p)$ matrix $N^{(p,p)} = \left(N_0^{(p)}, \ldots, N_p^{(p)} \right)$. Likewise, define $M_{jl}, M_j^{(p)}$ and a $r(p) \times r(p)$ matrix $M^{(p,p)}$ by replacing μ_k in the definitions of $N_{jl}, N_j^{(p)}$ and $N^{(p,p)}$ by $\kappa_k = \int u^k K^2(u)\, du$.

Now we let $\rho_{ij}(x) = \rho_{ij}(x, 0)$, and define $[r(p) + 1] \times [r(p) + 1]$ matrices

$$D(x) = \begin{bmatrix} N^{(p,p)} \rho_{20}(x) & N_0^{(p)} \rho_{11}(x) \\ N_0^{(p)T} \rho_{11}(x) & N_{00} \rho_{02}(x) \end{bmatrix}, \quad V(x) = \begin{bmatrix} M^{(p,p)} v_{20}(x) & M_0^{(p)} v_{11}(x) \\ M_0^{(p)T} v_{11}(x) & M_{00} v_{02}(x) \end{bmatrix},$$

where $v_{20}(x) = E\left[\ell_{10}^2(Y, f(x), \tau) \mid X = x \right]$, $v_{02}(x) = E\left[\ell_{01}^2(Y, f(x), \tau) \mid X = x \right]$, and $v_{11}(x) = E\left[\ell_{10}(Y, f(x), \tau)\ell_{01}(Y, f(x), \tau) \mid X = x \right]$. Note that if $\rho_{20}(x) > 0$, $\rho_{20}(x)\rho_{02}(x) - \rho_{11}(x)^2 > 0$

6

and $K \geq 0$ is supported on a set which contains a d-dimensional open rectangle, then $D(x)$ is positive definite. Under the latter condition on K, the matrix $V(x)$ is also positive definite unless there exists a nonzero constant c such that $\ell_{10}(Y, f(x), \tau) = c \, \ell_{01}(Y, f(x), \tau)$ with probability one, conditionally on $X = x$. Let $\mathcal{D}^{(p,p)}(x)$ be the $r(p) \times r(p)$ upper-left block of $\mathcal{D}(x) = D(x)^{-1}$.

To translate each of $\hat{\theta}_i(x)$ defined at (2.1) to an estimator of $f(x)$ or its derivatives, we consider blocks of size m_j, $j = 0, \ldots, p$, in the vector of $\hat{\theta}_i(x)$. Write $\hat{\theta}^{(0)}(x) = \hat{\theta}_0(x)$, and let $\hat{\theta}^{(j)}(x)$ for $j \geq 1$ be the jth block of size m_j defined by

$$\hat{\theta}^{(j)}(x) = [\hat{\theta}_{r(j-1)}(x), \ldots, \hat{\theta}_{r(j)-1}(x)]^T.$$

Thus, $\hat{\theta}^{(1)}(x) = \left[\hat{\theta}_1(x), \ldots, \hat{\theta}_d(x)\right]^T$, and so on. Furthermore,

$$\hat{\theta}(x) \equiv [\hat{\theta}_0(x), \ldots, \hat{\theta}_{r(p)-1}(x)]^T = [\hat{\theta}^{(0)}(x), \hat{\theta}^{(1)T}(x), \ldots, \hat{\theta}^{(p)T}(x)]^T.$$

Define $\mathcal{E}_j^{(p)}$ by $\mathcal{E}_j^{(p)T} = \left[O_{m_j \times r(j-1)}, I_{m_j}, O_{m_j \times (r(p)-r(j))}\right]$, where $O_{r \times s}$ denote the $r \times s$ zero matrix, and I_r is the r-dimensional identity matrix. For $j = 0, \ldots, p$, $\mathcal{E}_j^{(p)}$ is a $r(p) \times m_j$ matrix which maps the whole $\hat{\theta}(x)$ to $\hat{\theta}^{(j)}(x)$ by

$$\hat{\theta}^{(j)}(x) = \mathcal{E}_j^{(p)T} \hat{\theta}(x).$$

Let $\theta^{(j)}(x)$ be the m_j-vector of all the jth partial derivatives of $f(x)$ divided by the corresponding factorials, arranged in the counter-lexicographical order, i.e.,

$$\theta^{(j)}(x) = [\nabla^{\xi_j(1)} f(x)/\xi_j(1)!, \ldots, \nabla^{\xi_j(m_j)} f(x)/\xi_j(m_j)!]^T. \tag{3.2}$$

Then, $\hat{\theta}^{(j)}(x)$ is the local polynomial estimator of $\theta^{(j)}(x)$.

Let $g(x)$ denote the marginal density function of X. Let $\mathcal{U}_{j,f}^{(p:0)}$ be the $[r(p)+1] \times m_j$ matrix obtained by adding the row vector $O_{1 \times m_j}$ at the bottom of $\mathcal{E}_j^{(p)}$, i.e., $\mathcal{U}_{j,f}^{(p:0)T} = (\mathcal{E}_j^{(p)T}, O_{m_j \times 1})$. Also, we define

$$\ell^*(y, x) = \ell_{01}(y, f(x), \tau) + \rho_{11}(x)\mathcal{E}_0^{(p)T} \mathcal{D}(x) \left[\begin{array}{c} N_0^{(p)} \ell_{10}(y, f(x), \tau) \\ N_{00} \ell_{01}(y, f(x), \tau) \end{array} \right].$$

We obtain the following theorem.

Theorem 3.1. *Under the assumptions (A1)–(A9) given in Section 7, it follows that for each $j = 0, \ldots, p$*

$$\sqrt{nh^{2j+d}} \left[\hat{\theta}^{(j)}(x) - \theta^{(j)}(x) - h^{p-j+1}\rho_{20}(x)\mathcal{E}_j^{(p)T} \mathcal{D}^{(p,p)}(x) N_{p+1}^{(p)} \theta^{(p+1)}(x) + o(h^{p-j+1}) \right]$$

$$\xrightarrow{d} \mathcal{N}_{m_j} \left[0, \mathcal{U}_{j,f}^{(p:0)T} \mathcal{D}(x) V(x) \mathcal{D}(x) \mathcal{U}_{j,f}^{(p:0)}/g(x) \right],$$

7

where $\mathcal{N}_r$ denotes an r-variate normal distribution. For the estimator $\hat{\tau}$, we have under the assumptions (B1)–(B9)

$$\sqrt{n}\,(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}_1\left[0, \{E\rho_{02}(X)\}^{-2} E\ell^*(Y,X)^2\right].$$

The theorem tells that the pth order local polynomial estimators of the jth partial derivatives converge to the true values at the rate $h^{p-j+1} + n^{-1/2}h^{-d/2-j}$. In fact, if all the odd order moments of K vanish, i.e., $\int u^k K(u)\,du = 0$ for all d-tuples k with $|k| = 1, 3, \ldots$, and $p - j$ is even, then it can be shown that the leading bias term of order h^{p-j+1} is zero. In this case, if $\theta^{(p+2)}(x)$ as defined at (3.2) exists and is continuous, then the bias is of order h^{p-j+2}.

We note that $\hat{\tau}$ is a two-stage estimator with $\hat{f}$ at the first stage. To make negligible the bias due to the first stage estimation, one needs undersmoothing for estimation of f such that $nh^{2(p+1)} \to 0$, see the condition (B9) in Section 7.1. Recall that the optimal order of the bandwidth h for estimating f equals $n^{-1/\{2(p+1)+d\}}$. This kind of undersmoothing is typically required for two stage estimation to achieve $\sqrt{n}$-consistency, see Carroll et al. (1997), for example. It can be shown that the rate of convergence for $\tilde{\tau}(x)$ as an estimator of the constant τ equals $n^{-1/2}h^{-d/2}$, which is inferior to $n^{-1/2}$ that is achieved by τ. This is because $\tilde{\tau}(x)$ takes only a fraction of data of size nh^d as is seen in the local fitting procedure at (2.1). To the contrary, $\hat{\tau}$ uses the full likelihood with f being replaced by its estimator $\hat{f}$.

Theorem 3.1 is valid even when the error distribution is misspecified, as long as the equation $Q(a,b) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}$. In fact, we investigated the model misspecification issue in our simulation study presented in Section 4. We will consider misspecification for the error distribution, considering fat tail symmetric and skewed distributions for the process generating the data, whereas the likelihood will be build under the normal assumption for G.

3.2 Functional Shape Parameter Case

In this subsection we present the asymptotic distributions of $\hat{\theta}_j(x)$ for $j = 0, 1, \ldots, r$, and $\hat{\tau}_j(x)$ for $j = 0, 1, \ldots, s$, defined at (2.3). We slightly modify the definitions of the terms that are used in Subsection 3.1, whenever necessary, and introduce more to treat the case where the shape parameter τ is an unknown function.

With slight abuse of notation, we continue to use the same notation ρ_{ij} and v_{ij}, which

are now defined as

$$
\begin{aligned}
\rho_{ij}(x) &= -E\left[\ell_{ij}(Y, f(x), \tau(x)) \mid X = x\right], \\
v_{20}(x) &= E\left[\ell_{10}^2(Y, f(x), \tau(x)) \mid X = x\right], \\
v_{02}(x) &= E\left[\ell_{01}^2(Y, f(x), \tau(x)) \mid X = x\right], \\
v_{11}(x) &= E\left[\ell_{10}(Y, f(x), \tau(x))\ell_{01}(Y, f(x), \tau(x)) \mid X = x\right].
\end{aligned}
$$

For $r(p)$-vectors $a \equiv (a_0, \ldots, a_{r(p)-1})^T$ and $b \equiv (b_0, \ldots, b_{r(p)-1})^T$, we also modify the definitions of $Q_j(a, b)$ as

$$
\begin{aligned}
Q_1(a, b) &= \int E\Big[\ell_{10}\big(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1}u_d^p, \\
&\qquad\qquad \tau(x) + b_0 + b_1 u_1 + \cdots + b_{r(q)-1}u_d^q\big)\Big| X = x\Big] z_p(u)K(u)\, du, \\
Q_2(a, b) &= \int E\Big[\ell_{01}\big(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1}u_d^p, \\
&\qquad\qquad \tau(x) + b_0 + b_1 u_1 + \cdots + b_{r(q)-1}u_d^q\big)\Big| X = x\Big] z_q(u)K(u)\, du.
\end{aligned}
$$

As in the case of constant shape parameter, it can be also shown that, if (3.1) holds with $\rho_{ij}(x, \alpha) = -E\left[\ell_{ij}(Y, f(x) + \alpha_1, \tau(x) + \alpha_2) \mid X = x\right]$, and $K \geq 0$ is supported on a set which contains a d-dimensional open rectangle, then the system of equations $Q_1(a, b) = 0$ and $Q_2(a, b) = 0$ has the unique solution.

To state an analogue of Theorem 3.1, we need further notation. Define an $r(p) \times r(q)$ matrix $N^{(p,q)} = \left(N_0^{(p)}, \ldots, N_q^{(p)}\right)$. Also, define $N^{(q,p)} = \left(N_0^{(q)}, \ldots, N_p^{(q)}\right)$ and $N^{(q,q)} = \left(N_0^{(q)}, \ldots, N_q^{(q)}\right)$, which are $r(q) \times r(p)$ and $r(q) \times r(q)$, respectively, matrices. Likewise, define $M^{(p,q)}$, $M^{(q,p)}$ and $M^{(q,q)}$ with μ_k being replaced by $\kappa_k = \int u^k K^2(u)\, du$. With these matrices, definitions of $D(x)$ and $V(x)$ are modified as

$$
D(x) = \left[\begin{array}{cc} N^{(p,p)}\rho_{20}(x) & N^{(p,q)}\rho_{11}(x) \\ N^{(q,p)}\rho_{11}(x) & N^{(q,q)}\rho_{02}(x) \end{array}\right], \quad V(x) = \left[\begin{array}{cc} M^{(p,p)}v_{20}(x) & M^{(p,q)}v_{11}(x) \\ M^{(q,p)}v_{11}(x) & M^{(q,q)}v_{02}(x) \end{array}\right].
$$

These are $[r(p) + r(q)] \times [r(p) + r(q)]$ matrices.

As in Subsection 3.1, let $\mathcal{D}^{(p,p)}(x)$ be the $r(p) \times r(p)$ upper-left block of $\mathcal{D}(x) = D(x)^{-1}$. Let $\mathcal{D}^{(p,q)}(x)$ be the $r(p) \times r(q)$ upper-right block, $\mathcal{D}^{(q,p)}(x)$ be the $r(q) \times r(p)$ lower-left block, and $\mathcal{D}^{(q,q)}(x)$ be the $r(q) \times r(q)$ lower-right block of $\mathcal{D}(x)$. Define $\tau^{(j)}(x)$ in the same way as $\theta^{(j)}(x)$ with f replaced by τ in (3.2). Also, define

$$
\hat{\tau}^{(j)}(x) = \mathcal{E}_j^{(q)T}\hat{\tau}(x).
$$

9

To express the biases of the estimators, define

$$
\begin{aligned}
B_1(x) &= \mathcal{D}(x)\left[\begin{array}{c} h^{p+1}\rho_{20}(x)N_{p+1}^{(p)}\theta^{(p+1)}(x) \\ h^{p+1}\rho_{11}(x)N_{p+1}^{(q)}\theta^{(p+1)}(x) \end{array}\right], \\
B_2(x) &= \mathcal{D}(x)\left[\begin{array}{c} h^{q+1}\rho_{11}(x)N_{q+1}^{(p)}\tau^{(q+1)}(x) \\ h^{q+1}\rho_{02}(x)N_{q+1}^{(q)}\tau^{(q+1)}(x) \end{array}\right], \\
B_{1,f}(x) &= \rho_{20}(x)\mathcal{D}^{(p,p)}N_{p+1}^{(p)}\theta^{(p+1)}(x) + \rho_{11}(x)\mathcal{D}^{(p,q)}(x)N_{p+1}^{(q)}\theta^{(p+1)}(x), \\
B_{2,f}(x) &= \rho_{11}(x)\mathcal{D}^{(p,p)}N_{q+1}^{(p)}\tau^{(q+1)}(x) + \rho_{02}(x)\mathcal{D}^{(p,q)}(x)N_{q+1}^{(q)}\tau^{(q+1)}(x), \\
B_{1,\tau}(x) &= \rho_{20}(x)\mathcal{D}^{(q,p)}N_{p+1}^{(p)}\theta^{(p+1)}(x) + \rho_{11}(x)\mathcal{D}^{(q,q)}(x)N_{p+1}^{(q)}\theta^{(p+1)}(x), \\
B_{2,\tau}(x) &= \rho_{11}(x)\mathcal{D}^{(q,p)}N_{q+1}^{(p)}\tau^{(q+1)}(x) + \rho_{02}(x)\mathcal{D}^{(q,q)}(x)N_{q+1}^{(q)}\tau^{(q+1)}(x).
\end{aligned}
$$

Extending the definition of $\mathcal{U}_{j,f}^{(p:0)}$ in Subsection 3.1, let $\mathcal{U}_{j,f}^{(p:q)}$ be the $[r(p)+r(q)]\times m_j$ matrix obtained by adding the zero matrix $O_{r(q)\times m_j}$ at the bottom of $\mathcal{E}_j^{(p)}$, i.e., $\mathcal{U}_{j,f}^{(p:q)T} = (\mathcal{E}_j^{(p)T}, O_{m_j\times r(q)})$. Also, let $\mathcal{U}_{j,\tau}^{(p:q)}$ be the $[r(p)+r(q)]\times m_j$ matrix obtained by adding the zero matrix $O_{r(p)\times m_j}$ at the top of $\mathcal{E}_j^{(q)}$, i.e., $\mathcal{U}_{j,\tau}^{(p:q)T} = (O_{m_j\times r(p)}, \mathcal{E}_j^{(q)T})$. Define

$$
\hat{a}_j(x) = h^j\left\{\hat{\theta}^{(j)}(x) - \theta^{(j)}(x)\right\}, \quad \hat{b}_j(x) = h^j\left\{\hat{\tau}^{(j)}(x) - \tau^{(j)}(x)\right\},
$$

$\hat{a}(x) = \left(\hat{a}_1(x)^T, \ldots, \hat{a}_p(x)^T\right)^T$, and $\hat{b}(x) = \left(\hat{b}_1(x)^T, \ldots, \hat{b}_q(x)^T\right)^T$.

We obtain the following theorem.

Theorem 3.2. *Under the assumptions (B1)–(B9) given in Section 7, it follows that*

$$
\sqrt{nh^d}\left[\left(\begin{array}{c} \hat{a}(x) \\ \hat{b}(x) \end{array}\right) - B_1(x)h^{p+1} - B_2(x)h^{p+2} + o(h^{p+1} + h^{q+1})\right]
$$
$$
\xrightarrow{d} \mathcal{N}_{r(p)+r(q)}\left[0, \mathcal{D}(x)V(x)\mathcal{D}(x)/g(x)\right],
$$

where $\mathcal{N}_r$ denotes an r-variate normal distribution. Thus, for each $j = 0, \ldots, p$

$$
\sqrt{nh^{2j+d}}\left[\hat{\theta}^{(j)}(x) - \theta^{(j)}(x) - \mathcal{E}_j^{(p)T}(B_{1,f}h^{p-j+1} + B_{2,f}h^{q-j+1}) + o(h^{p-j+1} + h^{q-j+1})\right]
$$
$$
\xrightarrow{d} \mathcal{N}_{m_j}\left[0, \mathcal{U}_{j,f}^{(p:q)T}\mathcal{D}(x)V(x)\mathcal{D}(x)\mathcal{U}_{j,f}^{(p:q)}/g(x)\right],
$$

and, for each $j = 0, \ldots, q$,

$$
\sqrt{nh^{2j+d}}\left[\hat{\tau}^{(j)}(x) - \tau^{(j)}(x) - \mathcal{E}_j^{(q)T}(B_{1,\tau}h^{p-j+1} + B_{2,\tau}h^{q-j+1}) + o(h^{p-j+1} + h^{q-j+1})\right]
$$
$$
\xrightarrow{d} \mathcal{N}_{m_j}\left[0, \mathcal{U}_{j,\tau}^{(p:q)T}\mathcal{D}(x)V(x)\mathcal{D}(x)\mathcal{U}_{j,\tau}^{(p:q)}/g(x)\right].
$$

The theorem tells that both $\hat{\theta}^{(j)}$ and $\hat{\tau}^{(j)}$ have the same order of bias even if one fits locally polynomials of different degrees for f and τ. The leading biases for $\hat{\theta}^{(j)}$ and $\hat{\tau}^{(j)}$ are of the same order $h^{(p \wedge q)-j+1}$, where $p \wedge q = p$ if $p \leq q$ and $p \wedge q = q$ otherwise. Thus, the smaller of p and q determines the order of the bias for both $\hat{\theta}^{(j)}$ and $\hat{\tau}^{(j)}$. This means that fitting a lower order polynomial for τ may jeopardize the estimator of f. This is a new theoretical finding. It explains the failure of the local constant fit for τ in our preliminary experiment, and justifies consideration of higher-order local polynomial fit for τ even if one is interested in estimating the function f and its derivatives.

Here again, as discussed in Subsection 3.1, if all the odd moments of K vanish and $(p \wedge q) - j$ is even, then the leading bias terms of $\hat{\theta}^{(j)}$ and $\hat{\tau}^{(j)}$ are of order $h^{(p \wedge q)-j+2}$ provided that $\theta^{((p \wedge q)+2)}(x)$ and $\tau^{((p \wedge q)+2)}(x)$ exist and are continuous. Also, Theorem 3.2 is valid even when the error distribution is misspecified, as long as the equation $Q(a,b) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}^{r(q)}$. We investigated the model misspecification issue for the functional shape parameter case in our simulation study, by considering, here too, fat tail symmetric and skewed error distributions, when using the normal for building our estimator.

4 Simulation Results

While constructing scenarios we had in mind a dependent variable bounded between 1 and infinity, with distribution skewed towards the unity bound, with most observations falling in between 1 and 2. Intuitively, this would be an index (*e.g.*, the Debreu-Farrell efficiency index), whose reciprocal is then bounded between 0 and 100%, and most of which are between 50% and 100%. This is adequate to, for example, what many empirical studies report about production efficiencies of firms or countries (*e.g.*, see Kumar and Russell, 2002, Zelenyuk and Zheka, 2006, etc...). Such scenarios, with relatively small variation in the dependent variable and most of which being near the bound, are especially difficult to handle and so would be a good assessment experiment for our estimator.

Beyond the theoretical results of Section 3, the robustness of our estimator to misspecification errors is an empirical issue. So we will investigate how our estimator behaves under various misspecification of the error distribution. Our likelihood is computed under the local normal assumption and we will then also consider samples generated under a fat tails but symmetric distribution and under a skewed distribution. All other elements of the simulation scenario being kept the same. We will also consider a case where the true regression function is highly non-linear. We will mostly consider univariate cases which is useful for visually observing the scenario and performance of our estimator relative to the plot of the true model

and its traditional, fully parametric estimator. But we also show one case in a multivariate nonlinear heteroskedastic setup, showing how partial derivatives can be recovered.

In the Example 1 below (constant shape parameter) we use local constant model for τ and local linear fit for the regression f. In all the other cases, local linear approximation was used for the shape parameter τ and we will compare in each case the local linear and local quadratic fit for the regression function f. The latter is particularly useful when one is interested in estimating the derivatives of the regression function. In all the examples, we will be using the normal distribution for G to compute the likelihood function and the Gaussian kernel. The bandwidth was determined by cross validation in the lines of *e.g.* Kumbhakar *et al.* (2007). The true model in all the cases can be written as

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

with $\varepsilon_i \geq 1 - f(X_i)$. The true distribution G of the error term before truncation will be either a Normal, a Student with 5 d.f. or a Gamma with shape parameter $\alpha = 3$[1], and its conditional variance given $X_i = x$ is denoted by $\sigma_\varepsilon^2(x)$. In the examples that follow we chose specific values for the parameters and some particular heteroskedastic scenario, but we have also tried many different values for the parameters and for the heteroskedasticity and the results were qualitatively similar to those presented below and are not presented to save space.

4.1 Example 1. Linear Model with Homoskedasticity

In this scenario we assume homoskedastic variance before truncation, *i.e.*, $\sigma_\varepsilon(X_i) = \sigma$. So we mean here that the shape parameter of the error term is homoskedastic whereas, after truncation the model obviously is heteroskedastic, because the variance would depend on the truncation point $(1 - f(X_i))$. In addition, we assume also that the regression equation is linear:

$$f(X_i) = \beta_0 + \beta_1 X_i \tag{4.2}$$

In this case, the traditional truncated regression model with linear regression function, and with homoskedastic shape parameter (before truncation) would be based on fully parametric maximum likelihood methods and would provide efficient estimators of the parameters. This is the approach studied in Simar and Wilson (2007). The goal of doing this experiment is to investigate how much do we lose (in a univariate case) by using our semiparametric estimator described in Section 2.1 when the true error distribution is indeed Normal with

---

[1]To fix the notation here, $\alpha = 1$ would be the exponential case.

constant variance (so that the full parametric approach is correct) and how well our estimator resists to misspecification of the error distribution. Figure 1 visualizes this scenario and the estimation results for $\sigma = 0.3, \beta_0 = 1.5, \beta_1 = 0.5$. The X_i were generated according to a uniform $U(-2, 2)$ and the sample size $n = 200$.
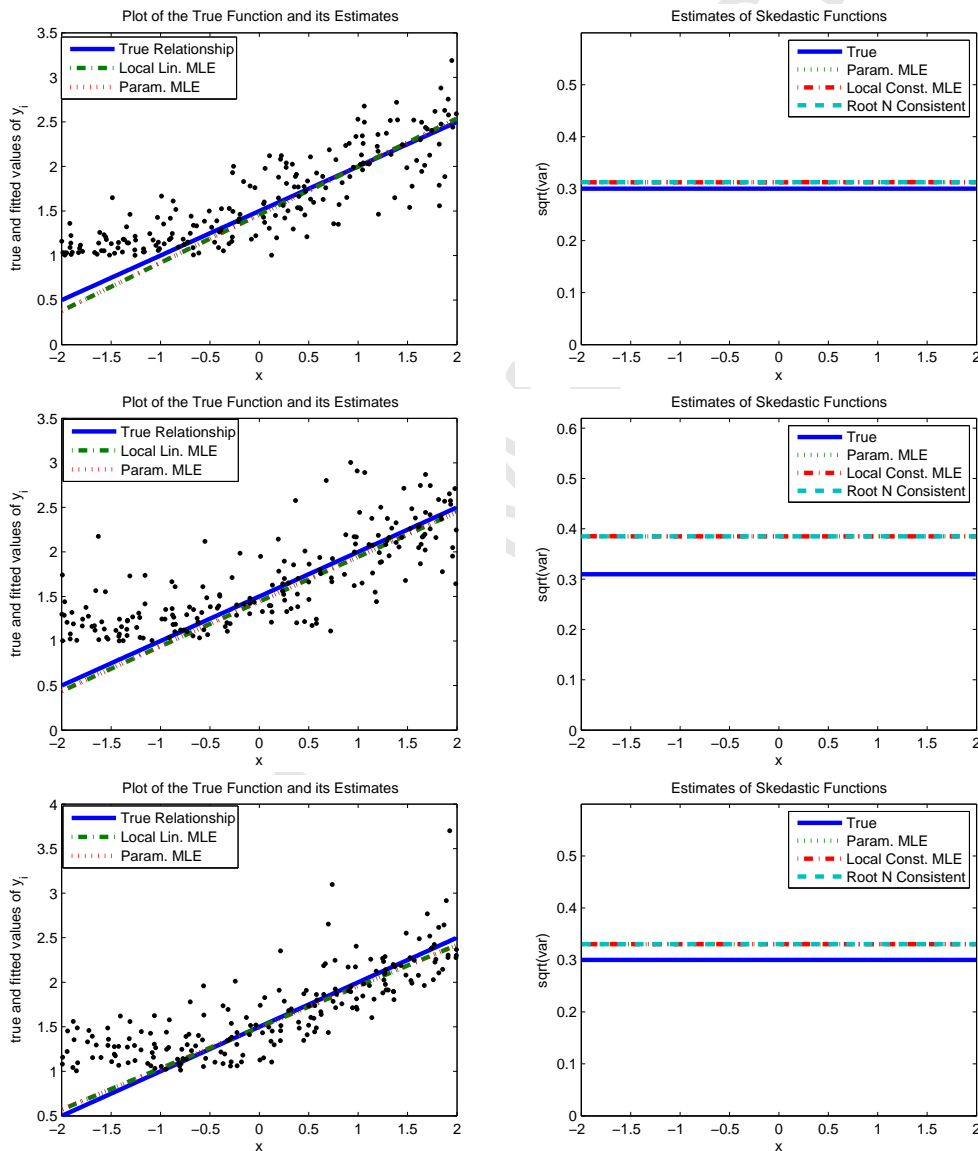

Figure 1: Example 1: Linear model with homoskedasticity. From top to bottom panels, Normal /Student/Gamma error terms.

Left panels of Figure 1 shows the plot of the true function (solid line) we want to estimate and the fit of two estimators: parametric ML estimator (dotted line) and local linear likelihood estimator (dash-dotted line). The right panels of the figure shows the plot of the

corresponding true and the three estimates of the constant shape parameter $\tau = \sigma_\varepsilon$ (the parametric MLE, the local constant $\tilde{\tau}(x)$ and the root-n semiparametric estimator $\hat{\tau}$ defined in Section 2.1). The error terms were simulated according (from top to bottom panels) truncated normal, student and gamma adjusted to provide the same true standard deviation $\sigma = 0.3$.

In the pictures we see that all the fits are very good (this was confirmed by other simulated samples with this scenario) and our semiparametric estimator coincides in all the cases to the parametric MLE, which is not a surprise since our anchorage model is the (true) linear model and so, the selected bandwidth is very large, pushing our local linear estimator to the simple parametric linear model. As a consequence, the estimation of τ are almost identical in all the cases and for all the approaches. It appears also that the approach is robust to the departures from normality of the error term analyzed here, but we observe slight overestimation of the variance in the Student and Gamma cases, especially in the Student case.

4.2 Example 2. Linear Model with Heteroskedasticity

In this scenario we assume heteroskedasticity for the shape parameter before truncation, i.e., $\sigma_\varepsilon(X_i)$ now depends on x. We present here only one illustrative case with $n = 400$ observations where $\sigma_\varepsilon(x) = 0.15\sqrt{3-x}$. Otherwise, the model is the same as in Example 1.

Here we compare also local linear (with a correct anchorage model) and local quadratic approximations for the regression. We use local linear approximation for the variance function. The results are provided in Figure 2. We also tried with a local constant approximation for the variance function and a local linear for the regression. The results are not reproduced for saving space. As expected they are very similar to those of the parametric MLE fit shown in Figure 2, since as in Example 1, the selected value of the bandwidth is very large and the semiparametric fit is almost identical to its anchorage linear model. This estimator persistently overestimated the regression function at the left end and, most importantly, estimated τ as a constant rather than a function of x, as also the fully parametric estimator did. This illustrates, from an empirical perspective, the usefulness of using at least linear approximation for the variance function, as suggested by the theory.

As Figure 2 shows, for the Normal case, the semiparametric estimators perform very well (both local linear and local quadratic), for both the estimation of the regression and of the variance function. In all the cases, the semiparametric estimator outperforms the standard parametric fit. For the Student and the Gamma cases, our semiparametric estimators resist well to departure from the normal even if the fit of the variance function is less accurate. In particular for the Gamma case, the local quadratic is less accurate than the local (true)

linear fit for the regression function. This might be due to the fact that the second order term essentially picks the heteroskedastic-skewness of the Gamma error term.
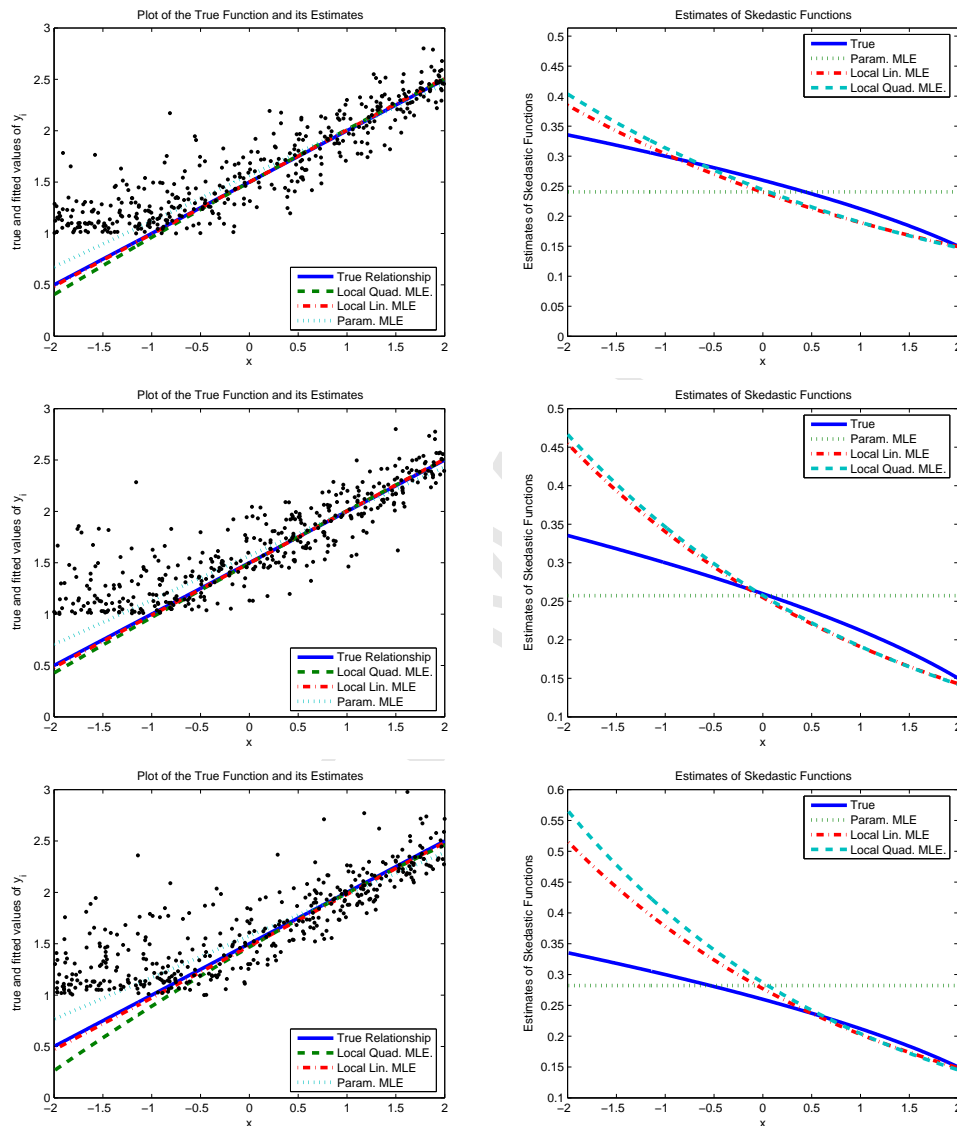

Figure 2: Example 2: Linear model with heteroskedasticity. From top to bottom panels, Normal /Student/Gamma error terms.

4.3 Example 3. Periodic Model with Heteroskedasticity

In this scenario we assume the regression function has some periodicity. In particular, we assume that

$$f(X_i) = \beta_0 + \beta_1 \sin(\gamma X_i) + \beta_2 X_i, \quad i = 1, \ldots, n. \tag{4.3}$$

Figure 3 shows a typical estimation result for $n = 400$. The specific values of the parameters in this example are $\beta_0 = 1, \beta_1 = 0.5, \beta_2 = 0.7, \gamma = \pi$ and $X_i \sim U(0,4)$. Moreover, to complicate the estimation problem, we also assume heteroskedasticity of σ_ε. In the illustration we chose $\sigma_\varepsilon(x) = 0.15\sqrt{6 - x}$, so that the closer to the truncation bound, the higher is the variance.
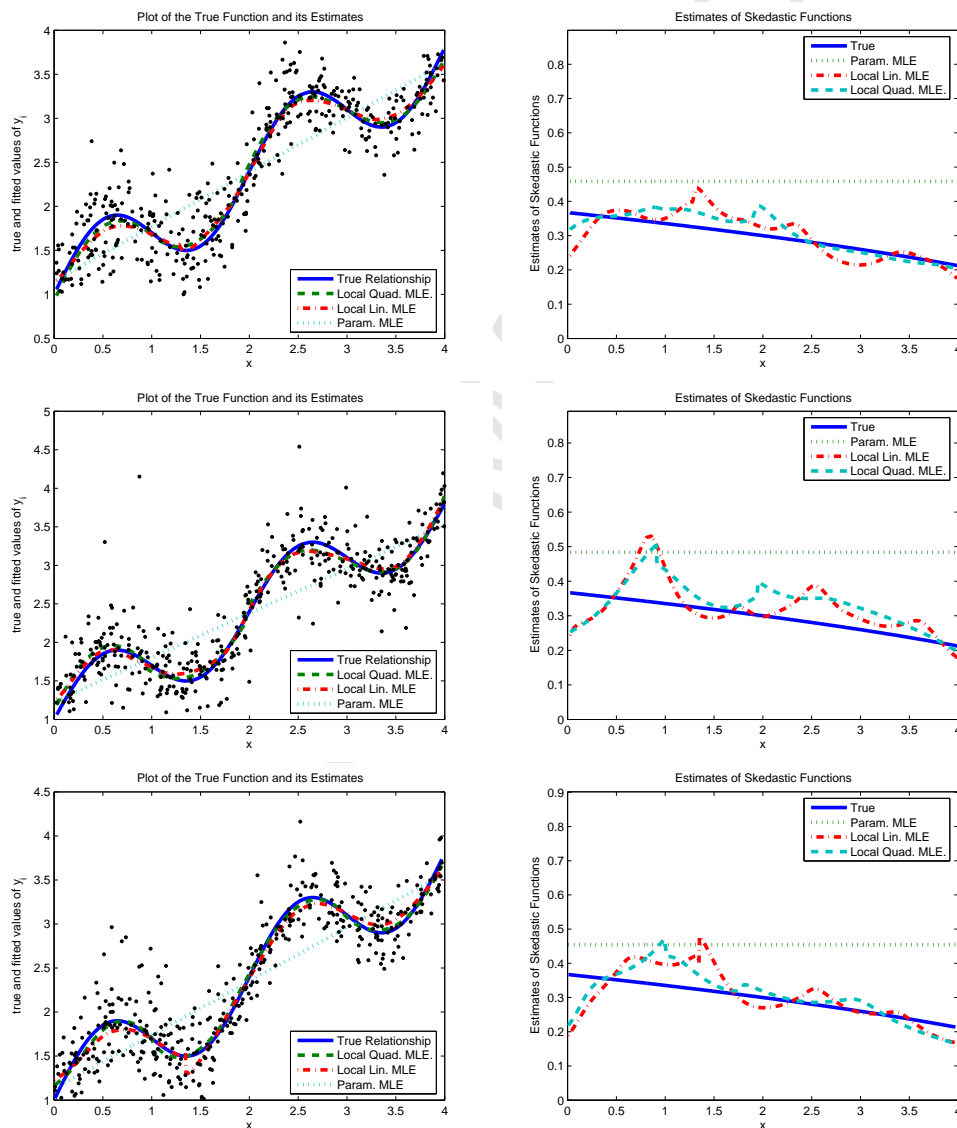

Figure 3: Example 3: Periodic model with heteroskedasticity. From top to bottom panels, Normal /Student/Gamma error terms.

The goal of the exercise is to see if our estimators perform well for relationships that seriously depart from linear or quadratic shape and, in fact, which might be fairly hard to guess

about in practice. The problem is also complicated with the presence of heteroskedasticity. Nevertheless, Figure 3 suggests that the performance is again quite good for both the linear and the quadratic estimators, with some robustness to the departure from normality of the error term, with again some overestimation of the variance function when the error is not normal.

4.4   Example 4. Multivariate Model with Heteroskedasticity

Here we consider two regressors that influence the dependent variable through a quadratic form. We want to see how our estimators perform for this type of scenario because the U-shape relationships are fairly common in economic phenomena. In addition, we want to see the performance when the situation is complicated by dependence of the variance on some of the regressors. For example, the employment level in a company may positively influence not only the mean of inefficiency but also company's risk (variance) of being inefficient, $e.g.$, because of increased risk of principal-agent problems, of pressure from trade unions, of strikes, etc... Formally, our scenario is given by (4.1), where for $i = 1, \ldots, n$:

$$f(X_i) = \beta_0 + \beta_{11}X_{1i} + \beta_{12}X_{1i}^2 + \beta_{21}X_{2i} + \beta_{22}X_{2i}^2 + \gamma X_{1i}X_{2i} \qquad (4.4)$$

with $\sigma_\varepsilon(X_i) = \sigma - \zeta(X_{1i} + \delta)^2$.

As before, we tried different values for the parameters and the results are quite robust. Figure 4 presents a typical result, where in this particular simulation we had $\beta_0 = 1.2, \beta_{11} = -0.1, \beta_{12} = 0.2, \beta_{21} = -0.1, \beta_{22} = 0.2, \gamma = -0.1, \sigma = 0.3, \zeta = 0.05, \delta = 1$ and $n = 200$. Note that for these particular values, heteroskedasticity is such that the variance increases near the truncation level, which complicates the estimation problem (homoskedastic case was also studied and good performance was also observed but to save space we do not present them). We see that the performance of the local linear is fairly good, but the quadratic one is much better.
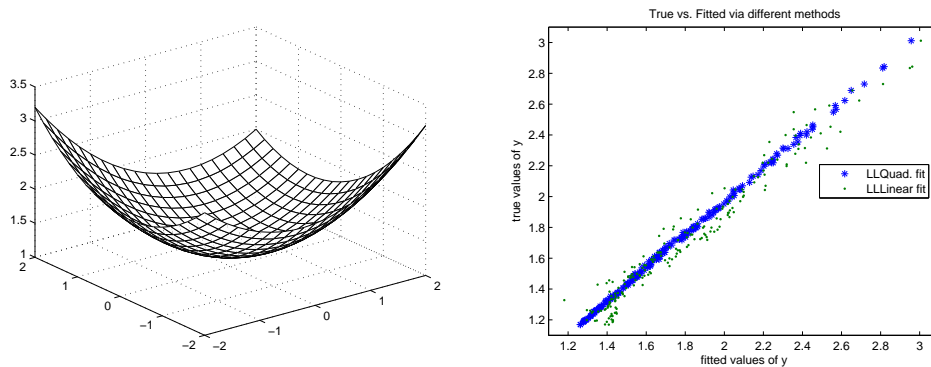
Figure 4: Example 5: Multivariate model with heteroskedastic shape parameter. Left panel is the model and right panel is the obtained fit.

Figure 5 gives a sense of the fit of the estimates of the partial derivatives of f w.r.t. X_1 and X_2, respectively. We see that the quadratic approximation substantially outperforms the linear one.
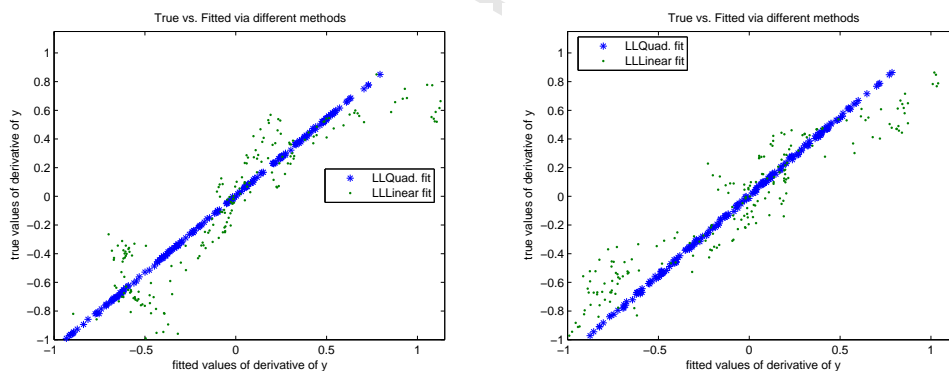

Figure 5: Example 5: Multivariate model with heteroskedastic shape parameter. Fit of the partial derivatives of f w.r.t. X_1 (left panel) and X_2 (right panel).

The observed superiority of the quadratic approach is not surprising at least for two reasons. First of all, the true model is quadratic and so it is natural that the local quadratic fit is better than the local linear one. Second, and more generally, the (finite-sample) bias of our estimator reduces with the order of local approximation of our estimator, as precisely stated in our Theorem 2.1. It is known that the higher the order approximations the better shall be the fit (*e.g.*, see Fan and Gijbels, 1996). In practice, however, researchers often stay satisfied with local linear estimators, motivating it with similar asymptotic properties but relative computational simplicity.

All our simulations suggest a different practical conclusion: One should definitely prefer the local quadratic likelihood estimator of the regression function relative to the local linear one, despite the increased computational complexity. This is especially true for the following cases: (i) when heteroskedasticity is expected; (ii) when one has many regressors with possible interaction among them; (iii) when the goal is to estimate the first partial derivatives of the regression function. And, these cases, are more the rules than exceptions in empirical studies. Higher order approximations (especially odd-order) theoretically should give better fit. However, even for third-order approximation, the programming cost and optimization cost increase dramatically and might not worth further improvements in the fit.

5   An Empirical Illustration

The goal of this section is not to make a solid empirical investigation but to get a feeling of the use and value of our estimator in studying economic phenomena. For this, we use data from a study about economic growth in the world, by Kumar and Russell (2002) that received considerable attention in the recent literature. Specifically, we take their estimated Farrell/Debreu-type efficiency scores for 57 countries in the world and relate it to capital-labor ratio (in the year 1990) in these countries[2].

We first use the same three estimators as in the simulations and obtain quite interesting results, presenting them in Figure 6. First of all, recall that the main argument of Kumar and Russell (2002) was that the change in capital per labor was the major source of economic growth in 1965-1990 and especially of the change from uni-modality to bi-modality of distribution of income per worker. The fully parametric linear model tells us that there is also negative (positive) relationship between the inefficiency (efficiency) of a country and its capital intensity. That is, the more capital per labor in a country the less inefficiency (the more efficiency) score of this country relative to the other countries. The estimated[3] slope parameter is 2.

---

[2]In the regression estimation we had to drop one observation (Switzerland) that appeared to be an outlier in terms of capital per worker and so causing computational problem in optimization of the likelihood function (even in fully parametric case).

[3]We used the Algorithm 1 of Simar and Wilson (2007) to obtain the estimates. Also note that part of the parametrically estimated curve is not observed in the left panel of Figure 6 because we trimmed the vertical axis.
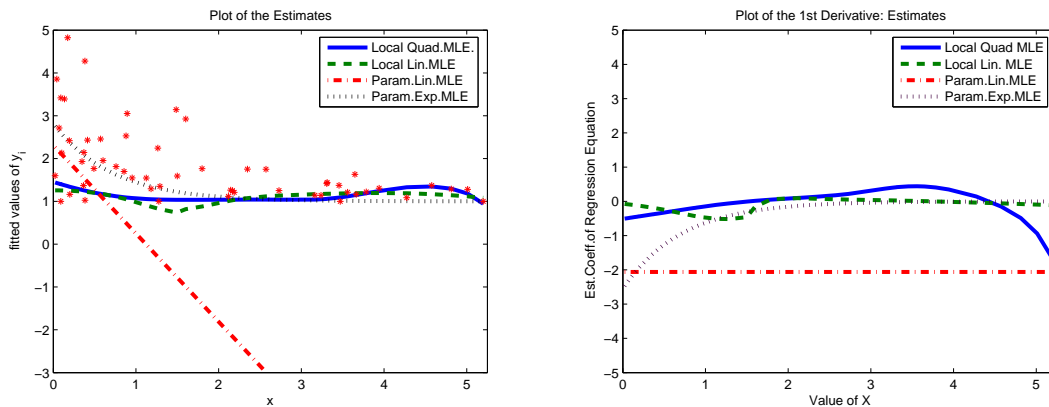
Figure 6: Empirical illustration. Left panel is the model and right panel is the derivative.

On the other hand, observing the left panel of Figure 6, one can see that heteroskedasticity is likely to be present in data: the less capital-labor the larger is the variance of inefficiency variable. So the apparent negative relationship can in fact be a result of severe heteroskedasticity.

From the plot of the linear fit in Figure 6, we might guess that the linear parametric model might be inappropriate here, and exponential might be a better choice. Of course, in practice such visually based conclusions on the parametric form can hardly be done for multivariate regressions, but this is useful for illustration and discussion here. We thus estimated the exponential (homoskedastic) model, $Y = 1 + \exp(X\beta) + \varepsilon$, which corresponds to the dotted curves in Figure 6. We see that the relationship between the capital depth and the inefficiency is indeed suggested to be negative, with relatively high marginal effect at the low capital per worker ratio and monotonically decreasing to almost no effect at the higher levels. We could also try various forms of heteroskedasticity with this or another functional form, but guessing about the two functional forms for the regression and for the shape parameter at the same time might be too much for a scientific approach.

So, instead, we try our non-parametric procedure that is capable of handling heteroskedasticity of unknown form and we get quite different conclusion than what the parametric models told us. Both linear and quadratic local likelihood estimators suggest that there is virtually no relationship between the Farrell/Debreu-type efficiency score of a country *vs.* capital-labor of this country. Specifically, the fitted curve characterizing the relationship is almost flat and the slope coefficient is fluctuating near zero. We see an exception at the very end (top 10% percentile) of the empirical range of the explanatory variable, where the quadratic approximation suggests that the relationship might indeed be negative, but this is only in a

small interval where there is only a few observations.

The results from estimating the regression equation non-parametrically makes us conjecture that the negative relationship between countries inefficiency score and its capital-labor ratio is coming not through the level (mean) of inefficiency but through the dispersion of inefficiency. Intuitively, we can say that the less capital-labor in a country the greater is the *risk* of having high inefficiency score for that country. Figure 7 gives a plot of the estimated variance of the error term *vs.* the explanatory variable, which supports our conjecture.
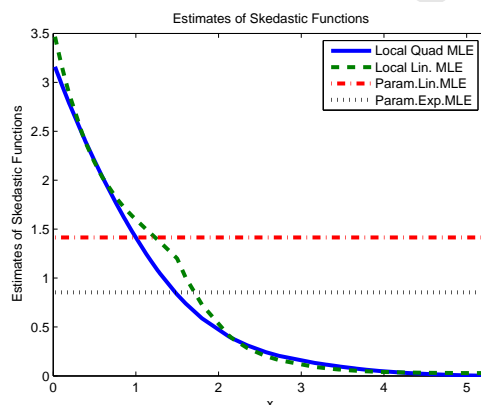

Figure 7: Empirical illustration. Estimation of the heterskedastic shape function.

The result we obtained in our small application is consistent with capitalist philosophy: if people possess a lot of capital within a country then they have a lot of incentives to create a political system that would minimize the risk of underutilization (inefficiency) of their capital. On the other hand, if people possess little of capital, "they have got little to lose", and so might be not as much interested or capable to build appropriate institutions that would protect property rights, ensure positive incentives for efficient allocation of resources and efforts and thus minimize the risk of underdevelopment. Looking at the data confirms that it is mostly the underdeveloped countries that are in the range of high variance of inefficiency scores and low capital per labor levels.

Finally, it might be worth reminding that in this section we had not intended to resolve the puzzles of economic growth and differences in efficiency across countries. Such study would require larger data set and more variables. Our goal was just to give a small illustration of our estimator in practice and show how it could suggest radically different conclusions than those obtained from commonly used, and sometimes abused, parametric methods.

6 Concluding Remarks

In this work we proposed a fairly flexible estimator for the context of truncated regression that does not require parametric assumptions. For this, we extended the theory of *local* maximum likelihood estimation, in particular the recent work of Kumbhakar *et al.* (2007), to the truncated case. We provided the asymptotic results of our estimator. Specifically, the estimator is consistent and asymptotically normally distributed with variance that can be estimated from data. Also, we derived a root-n consistent semiparametric estimator of the shape parameter in the homoskedastic case.

We also illustrated the performances of two variants of our estimator (namely, linear and quadratic approximations) on various simulated data sets, comparing it to the truth and to the parametric estimator. Remarkably, for the univariate case, our estimator performs as good as the traditional, fully parametric estimator when the assumptions for the latter hold, *i.e.*, we do not lose virtually anything by allowing the flexibility. However, our estimator performs much better when the parametric assumptions on the regression equation does not hold or even only when the assumption of homoskedastic shape parameter of the error term does not hold. We investigated how our semiparametric estimator behaved under departure from the normality assumption of the error term. For both "fat tails" departure (Student with 5 d.f.) and skewness (Gamma with shape parameter equal to 3) we have seen that our estimator is rather robust and still able to capture most of the structure of the underlying data generating process.

The use of our estimator on a real data set was illustrated from the recent study of Kumar and Russell (2002), analyzing relationship between the efficiency scores and the capital deepness in countries in the world. In this application we noticed that quite different and perhaps more plausible implications can be inferred using our estimator instead of the commonly used parametric one.

It became common that empirical researchers are often satisfied with local linear estimators, motivating it with similar asymptotic properties but relative computational simplicity. However, all our simulations suggested that, despite some increase in the computational complexity, the local quadratic likelihood estimator of the regression function should be preferred relative to the local linear one, especially if heteroskedasticity is expected and certainly when the focus is on estimating first derivatives. We also found that fitting a lower order polynomial for the shape parameter τ may jeopardize the estimator of the regression function f. This is a new theoretical finding, which was supported with preliminary simulation results and justifies consideration of higher-order local polynomial for τ even if one is interested in estimating the function f and its derivatives.

Noteworthy, our study can be extended in many ways. One natural extension would be to investigate endogeneity problem. Another extension is to incorporate the use of discrete explanatory variables. Yet another extension is to analyze the panel data framework. Our study is just the start, telling that the non-parametric estimator in the truncated regression context should exhibit a fair degree of robustness to various deviations from linearity in the regression equation and in the function defining the heteroskedastic shape parameter and to departure from the normality assumption. Thus our semiparametric estimator should prove to be very useful in practical applications.

7 Regularity Conditions and Proof of Theorems

7.1 Regularity Conditions

First, we collect the assumptions for the first part of Theorem 3.1. For $a \in \mathbb{R}^{r(p)}$ and $b \in \mathbb{R}$, let $\psi(a, b|x) \equiv [\psi_1(a, b|x), \psi_2(a, b|x)]^T$ where

$$
\begin{aligned}
\psi_1(a, b|x) &= E\Big[\ell_{10}\big(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1} u_d^p, \tau + b\big)\Big| X = x\Big], \\
\psi_2(a, b|x) &= E\Big[\ell_{01}\big(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1} u_d^p, \tau + b\big)\Big| X = x\Big].
\end{aligned}
$$

(A1) $Q(a, b) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}$;

(A2) $\sup_{(a^T, b)^T \in \mathcal{A}} \left| \psi(a, b|x + z) - \psi(a, b|x) \right| \to 0$ as $z \to 0$ for some compact set $\mathcal{A} \subset \mathbb{R}^{r(p)+1}$;

(A3) for $(i, j) = (1, 0)$ and $(0, 1)$, the following condition holds: for any compact sets $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}$, there exist functions U_{ij} such that $\sup_{\nu \in \mathcal{A}_1, \omega \in \mathcal{A}_2} |\ell_{ij}(y, \nu, \omega)| \le U_{ij}(y)$ and $\sup_{|z-x| \le \varepsilon} E\big(U_{ij}^{2+\delta}(Y)|X = z\big) < \infty$ for some $\varepsilon, \delta > 0$;

(A4) for $(i, j) = (2, 0)$, $(0, 2)$ and $(1, 1)$, the following condition holds: $\ell_{ij}(y, \nu, \omega)$ are continuous in (ν, ω) for each y, and for any compact sets $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}$, there exist functions U_{ij} such that $\sup_{\nu \in \mathcal{A}_1, \omega \in \mathcal{A}_2} |\ell_{ij}(y, \nu, \omega)| \le U_{ij}(y)$ and $\sup_{|z-x| \le \varepsilon} E\big(U_{ij}^2(Y)|X = z\big) < \infty$ for some $\varepsilon > 0$;

(A5) $g(x) > 0$, $\rho_{20}(x) > 0$, $\rho_{20}(x)\rho_{02}(x) - \rho_{11}(x)^2 > 0$, $v_{20}(x) > 0$, $v_{20}(x)v_{02}(x) - v_{11}(x)^2 > 0$

(A6) g, all ρ_{ij} and v_{ij} for $(i, j) = (2, 0), (0, 2), (1, 1)$ are continuous at x;

(A7) K is nonnegative, bounded and supported on $[-1, 1]^d$;

(A8) The function f has $(p+1)$th continuous partial derivatives at x;

(A9) $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and $nh^{2p+d+2} < C$ for some positive constant C.

For the second part of Theorem 3.1, we need the conditions for the first part hold uniformly for x in its support $\mathcal{S}$. We assume $\mathcal{S}$ is compact. For the statement of the conditions, we write $Q(a, b; x)$ for $Q(a, b)$ defined in Section 3.1 to stress its dependence on x.

(B1) $Q(a, b; x) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}$ for all $x \in \mathcal{S}$;

(B2) $\sup_{(a^T, b)^T \in \mathcal{A}, x \in \mathcal{S}} \left| \psi(a, b|x + z) - \psi(a, b|x) \right| \to 0$ as $z \to 0$ for some compact set $\mathcal{A} \subset \mathbb{R}^{r(p)+1}$;

(B3) for $(i, j) = (1, 0)$ and $(0, 1)$, the following condition holds: for any compact sets $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}$, there exist functions U_{ij} such that $\sup_{\nu \in \mathcal{A}_1, \omega \in \mathcal{A}_2} |\ell_{ij}(y, \nu, \omega)| \le U_{ij}(y)$ and $\sup_{x \in \mathcal{S}} E\left(U_{ij}^{2+\delta}(Y)|X = x\right) < \infty$ for some $\delta > 0$;

(B4) for $(i, j) = (2, 0)$, $(0, 2)$ and $(1, 1)$, the following condition holds: $\ell_{ij}(y, \nu, \omega)$ are continuous in (ν, ω) for each y, and for any compact sets $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}$, there exist functions U_{ij} such that $\sup_{\nu \in \mathcal{A}_1, \omega \in \mathcal{A}_2} |\ell_{ij}(y, \nu, \omega)| \le U_{ij}(y)$ and $\sup_{x \in \mathcal{S}} E\left(U_{ij}^2(Y)|X = x\right) < \infty$;

(B5) $\inf_{x \in \mathcal{S}} g(x) > 0$, $\inf_{x \in \mathcal{S}} \rho_{20}(x) > 0$, $\inf_{x \in \mathcal{S}}\{\rho_{20}(x)\rho_{02}(x) - \rho_{11}(x)^2\} > 0$, $\inf_{x \in \mathcal{S}} v_{20}(x) > 0$, $\inf_{x \in \mathcal{S}}\{v_{20}(x)v_{02}(x) - v_{11}(x)^2\} > 0$

(B6) g, all ρ_{ij} and v_{ij} for $(i, j) = (2, 0)$, $(0, 2)$, $(1, 1)$ are continuous on $\mathcal{S}$;

(B7) K is nonnegative, bounded and supported on $[-1, 1]^d$;

(B8) The function f has $(p+1)$th continuous partial derivatives on $\mathcal{S}$;

(B9) $nh^{2(p+1)} \to 0$ and $nh^{2d}/(\log n)^2 \to \infty$ as $n \to \infty$.

Now, to list the assumptions for Theorem 3.2, let $\psi(a, b|x) \equiv [\psi_1(a, b|x), \psi_2(a, b|x)]^T$ for $a \in \mathbb{R}^{r(p)}$ and $b \in \mathbb{R}^{r(q)}$, where

$$\psi_1(a, b|x) = E\Big[\ell_{10}\big(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1}u_d^p,$$
$$\tau(x) + b_0 + b_1 u_1 + \cdots + b_{r(q)-1}u_d^q\big)\Big| X = x\Big],$$
$$\psi_2(a, b|x) = E\Big[\ell_{01}\big(Y, f(x) + a_0 + a_1 u_1 + \cdots + a_{r(p)-1}u_d^p,$$
$$\tau(x) + b_0 + b_1 u_1 + \cdots + b_{r(q)-1}u_d^q\big)\Big| X = x\Big].$$

24

(C1) $Q(a, b) = 0$ has the unique solution $a = 0 \in \mathbb{R}^{r(p)}$ and $b = 0 \in \mathbb{R}^{r(q)}$;

(C2) $\sup_{(a^T, b^T)^T \in \mathcal{A}} \left| \psi(a, b|x + z) - \psi(a, b|x) \right| \to 0$ as $z \to 0$ for some compact set $\mathcal{A} \subset \mathbb{R}^{r(p) + r(q)}$;

(C3)–(C7) same as (A3)–(A7);

(C8) The function f has $(p + 1)$th continuous partial derivatives, and the function τ has $(q + 1)$th continuous partial derivatives, at x;

(C9) $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and $nh^{2(p \wedge q) + d + 2} < C$ for some positive constant C.

7.2 Proofs of Theorems 3.1 and 3.2

We outline a proof of Theorem 3.2 first, and then give a proof of Theorem 3.1.

7.2.1 Proof of Theorem 3.2

Define $u^{(j)} = \left(u^{\xi_j(1)}, \ldots, u^{\xi_j(m_j)} \right)^T$ for a d-vector u, where $\xi_j(s)$ is defined in Section 3. Let $\tilde{f}$ be the pth order polynomial approximation of f around the point x, and $\tilde{\tau}$ the qth order polynomial approximation of τ, i.e.,

$$
\begin{aligned}
\tilde{f}(u) &= \sum_{j=0}^{p} \theta^{(j)}(x)^T (u - x)^{(j)}, \\
\tilde{\tau}(u) &= \sum_{j=0}^{q} \tau^{(j)}(x)^T (u - x)^{(j)},
\end{aligned}
$$

where $\theta^{(j)}(x)$ is defined at (3.2). Define

$$
\begin{aligned}
\hat{a}^{(j)} &\equiv \hat{a}^{(j)}(x) = h^j \left(\hat{\theta}^{(j)}(x) - \theta^{(j)} \right), \; j = 0, \ldots, p, \\
\hat{b}^{(j)} &\equiv \hat{b}^{(j)}(x) = h^j \left(\hat{\tau}^{(j)}(x) - \tau^{(j)} \right), \; j = 0, \ldots, q.
\end{aligned}
$$

25

Also, define

$$
Z_{p,i} = \left(1, \left(\frac{X_{i1} - x_1}{h}\right), \ldots, \left(\frac{X_{id} - x_d}{h}\right)^p\right)^T,
$$

$$
\tilde{\ell}_{jk}(i, a, b) = \ell_{jk}\left(Y_i, \tilde{f}(X_i) + a_0 + a_1\left(\frac{X_{i1} - x_1}{h}\right) + \cdots + a_{r(p)-1}\left(\frac{X_{id} - x_d}{h}\right)^p,\right.
$$

$$
\left. \tilde{\tau}(X_i) + b_0 + b_1\left(\frac{X_{i1} - x_1}{h}\right) + \cdots + b_{r(q)-1}\left(\frac{X_{id} - x_d}{h}\right)^q\right)
$$

$$
Q_{1n}(a, b) = n^{-1}\sum_{i=1}^{n} Z_{p,i}\tilde{\ell}_{10}(i, a, b)K_h(X_i - x),
$$

$$
Q_{2n}(a, b) = n^{-1}\sum_{i=1}^{n} Z_{q,i}\tilde{\ell}_{01}(i, a, b)K_h(X_i - x).
$$

Write $\hat{a} = (\hat{a}^{(0)T}, \ldots, \hat{a}^{(p)T})^T$ and $\hat{b} = (\hat{b}^{(0)T}, \ldots, \hat{b}^{(q)T})^T$. Then $(\hat{a}, \hat{b})$ is the solution of the equation $Q_n(a, b) = 0$, where $Q_n(a, b) = \left(Q_{1n}(a, b)^T, Q_{2n}(a, b)^T\right)^T$ is a $[r(p) + r(q)]$-vector. One can prove in a similar way as in Kumbhakar *et al.* (2007) that for any compact set $\mathcal{A}$

$$
\sup_{(a,b)\in\mathcal{A}} \left|Q_n(a, b) - EQ_n(a, b)\right| = O_p\left(n^{-1/2}h^{-d/2}(\log n)^{1/2}\right), \tag{7.1}
$$

$$
\sup_{(a,b)\in\mathcal{A}} \left|EQ_n(a, b) - Q(a, b)\right| = o(1). \tag{7.2}
$$

By (7.1), (7.2) and the assumption (C1) it follows that

$$
\hat{a} = o_p(1), \quad \hat{b} = o_p(1). \tag{7.3}
$$

Next, let $S_n(a, b)$ be a $[r(p) + r(q)] \times [r(p) + r(q)]$ matrix defined as

$$
S_n(a, b) = n^{-1}\sum_{i=1}^{n}\left[\begin{array}{cc} Z_{p,i}Z_{p,i}^T\tilde{\ell}_{20}(i, a, b), & Z_{p,i}Z_{q,i}^T\tilde{\ell}_{11}(i, a, b) \\[2mm] Z_{q,i}Z_{p,i}^T\tilde{\ell}_{11}(i, a, b), & Z_{q,i}Z_{q,i}^T\tilde{\ell}_{02}(i, a, b) \end{array}\right]K_h(X_i - x)
$$

It can be proved that for any compact set $\mathcal{A}$

$$
\sup_{(a,b)\in\mathcal{A}} \left|S_n(a, b) - ES_n(a, b)\right| = O_p\left(n^{-1/2}h^{-d/2}(\log n)^{1/2}\right) \tag{7.4}
$$

A Taylor expansion of $Q_n(a, b)$ gives

$$
0 = Q_n(\hat{a}, \hat{b}) = Q_n(0, 0) + S_n(a^*, b^*)\left[\begin{array}{c} \hat{a} \\ \hat{b} \end{array}\right], \tag{7.5}
$$

where a^* and b^* are random vectors such that $|(a^{*T}, b^{*T})^T| \leq |(\hat{a}^T, \hat{b}^T)^T|$. The consistency of $(\hat{a}, \hat{b})$ as given at (7.3) and the result (7.4) imply

$$
S_n(a^*, b^*) - ES_n(0, 0) = o_p(1). \tag{7.6}
$$

26

Furthermore, one can show that

$$ES_n(0,0) = -D(x)g(x) + o(1). \tag{7.7}$$

By (7.5)–(7.7) and the fact that $Q_n(0,0) = O_p(n^{-1/2}h^{-d/2})$, we obtain the following expansion for $(\hat{a}^T, \hat{b}^T)^T$:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = g(x)^{-1}D(x)^{-1}Q_n(0,0) + o_p(n^{-1/2}h^{-d/2}). \tag{7.8}$$

The mean and variance of $(\hat{a}^T, \hat{b}^T)^T$ come from those of $Q_n(0,0)$. One can show

$$E[Q_n(0,0)] = \begin{bmatrix} h^{p+1}\rho_{20}(x)N_{p+1}^{(p)}\theta^{(p+1)}(x) + h^{q+1}\rho_{11}(x)N_{q+1}^{(p)}\tau^{(q+1)}(x) \\ h^{p+1}\rho_{11}(x)N_{p+1}^{(q)}\theta^{(p+1)}(x) + h^{q+1}\rho_{02}(x)N_{q+1}^{(q)}\tau^{(q+1)}(x) \end{bmatrix} g(x) \tag{7.9}$$
$$+ o(h^{p+1} + h^{q+1}),$$

$$\mathrm{var}[Q_n(0,0)] = n^{-1}h^{-d} \begin{bmatrix} M^{(p,p)}v_{20}(x) & M^{(p,q)}v_{11}(x) \\ M^{(q,p)}v_{11}(x) & M^{(q,q)}v_{02}(x) \end{bmatrix} g(x) + o(n^{-1}h^{-d}). \tag{7.10}$$

The asymptotic normality of $Q_n(0,0)$ follows from the assumption (C3) by a direct application of the Lindeberg-Feller theorem. The theorem now follows from the asymptotic normality of $Q_n(0,0)$ and (7.8)–(7.10).

7.2.2   Proof of Theorem 3.1

With slight abuse of notation, we continue to use $\tilde{\ell}_{jk}(i,a,b)$ to denote

$$\tilde{\ell}_{jk}(i,a,b) = \ell_{jk}\left(Y_i, \tilde{f}(X_i) + a_0 + a_1\left(\frac{X_{i1}-x_1}{h}\right) + \cdots + a_{r(p)-1}\left(\frac{X_{id}-x_d}{h}\right)^p, \tau + b\right),$$

where b is now a scalar. Also, we define $Q_{jn}(a,b)$ for $j = 1, 2$ and $S_n(a,b)$ as in the proof of Theorem 3.2, but with replacing $Z_{q,i}$ by 1. Then, we get (7.8) where $D(x)$ is now as defined in Section 3.1. Since in this case $\tau^{(j)} = 0$ for all $j \geq 1$, we obtain instead of (7.9) and (7.10), respectively,

$$E[Q_n(0,0)] = \begin{bmatrix} h^{p+1}\rho_{20}(x)N_{p+1}^{(p)}\theta^{(p+1)}(x) \\ h^{p+1}\rho_{11}(x)N_{0,p+1}\theta^{(p+1)}(x) \end{bmatrix} g(x) + o(h^{p+1}), \tag{7.11}$$

$$\mathrm{var}[Q_n(0,0)] = n^{-1}h^{-d}V(x)g(x) + o(n^{-1}h^{-d}),$$

where $V(x)$ is as defined in Section 3.1. The first part of the theorem follows immediately from these expansions.

For the proof of the second part, we need a stronger result than (7.8). To do this, define $(\overline{a}(x)^T, \overline{b}(x))^T = -\{ES_n(0,0;x)\}^{-1}Q_n(0,0;x)$. Note that here we make explicit the dependence of $Q_n(0,0)$ and $S_n(0,0)$ on x. Also, we write $\tilde{f}(u;x)$ and $Z_{p,i}(x)$ instead of $\tilde{f}(u)$

27

and $Z_{p,i}$ defined in the proof of Theorem 3.2. Let $c_n = n^{-1/2}h^{-d/2}(\log n)^{1/2}$. From (7.11) and the fact that $\sup_{x\in\mathcal{S}}|Q_n(0,0;x) - EQ_n(0,0;x)| = O_p(c_n)$, we see that

$$Q_n(0,0;x) = O_p(c_n + h^{p+1}) \tag{7.12}$$

uniformly for x in $\mathcal{S}$. Now, let $\Delta_1(x) = \hat{a}(x) - \overline{a}(x)$ and $\Delta_2(x) = \hat{b}(x) - \overline{b}(x)$. Write $\Delta(x) = \left(\Delta_1(x)^T, \Delta_2(x)\right)^T$. Then, from the expansion

$$
\begin{aligned}
0 = Q_n(\hat{a}(x), \hat{b}(x)) &= Q_n(\overline{a}(x), \overline{b}(x); x) + S_n(\overline{a}(x), \overline{b}(x); x)\Delta(x)\{1 + o_p(1)\} \\
&= Q_n(\overline{a}(x), \overline{b}(x); x) + S_n(0,0;x)\Delta(x)\{1 + o_p(1)\},
\end{aligned}
$$

we obtain

$$\Delta(x) = -S_n(0,0;x)^{-1}Q_n(\overline{a}(x), \overline{b}(x); x)\{1 + o_p(1)\} \tag{7.13}$$

uniformly for x in $\mathcal{S}$. Furthermore, we have from (7.12) that, uniformly for $x \in \mathcal{S}$,

$$
\begin{aligned}
Q_n(\overline{a}(x), \overline{b}(x); x) &= Q_n(0,0;x) + S_n(0,0;x)\left[\begin{array}{c} \overline{a}(x) \\ \overline{b}(x) \end{array}\right] + O_p(c_n^2 + h^{2(p+1)}) \\
&= Q_n(0,0;x) - S_n(0,0;x)\{ES_n(0,0;x)\}^{-1}Q_n(0,0;x) + O_p(c_n^2 + h^{2(p+1)}) \\
&= -\{S_n(0,0;x) - ES_n(0,0;x)\}\{ES_n(0,0;x)\}^{-1}Q_n(0,0;x) \\
&\quad + O_p(c_n^2 + h^{2(p+1)}) \\
&= O_p(c_n^2 + h^{2(p+1)}).
\end{aligned}
$$

This and (7.13) give $\sup_{x\in\mathcal{S}}|\Delta(x)| = O_p(c_n^2 + h^{2(p+1)})$. Thus, we obtain for the estimator $\hat{f}(x)$ of $f(x) = \theta^{(0)}(x)$

$$\sup_{x\in\mathcal{S}}\left|\hat{f}(x) - f(x) - n^{-1}\sum_{i=1}^{n}W_i(x)K_h(X_i - x)\right| = O_p(c_n^2 + h^{2(p+1)}), \tag{7.14}$$

where $W_i(x)$ is the first element of the vector,

$$H_n(x)^{-1}\left[\begin{array}{c} Z_{p,i}(x)\tilde{\ell}_{10}(i,0,0) \\ \tilde{\ell}_{01}(i,0,0) \end{array}\right] = H_n(x)^{-1}\left[\begin{array}{c} Z_{p,i}(x)\ell_{10}(Y_i, \tilde{f}(X_i;x), \tau) \\ \ell_{01}(Y_i, \tilde{f}(X_i;x), \tau) \end{array}\right]$$

and $H_n(x) = -\{ES_n(0,0;x)\}$.

Define $\hat{d} = \hat{\tau} - \tau$. Then, $\hat{d}$ is the solution of the equation $Q_{3n}(d) = 0$, where

$$Q_{3n}(d) = n^{-1}\sum_{i=1}^{n}\ell_{01}(Y_i, \hat{f}(X_i), \tau + d).$$

Replacing $\hat{f}$ by f in $Q_{3n}(d)$, let

$$Q_{4n}(d) = n^{-1}\sum_{i=1}^{n}\ell_{01}(Y_i, f(X_i), \tau + d).$$

28

It follows that

$$
\begin{aligned}
Q_{3n}(d) &= Q_{4n}(d) + n^{-1}\sum_{i=1}^{n}\ell_{11}(Y_i, f(X_i), \tau + d)\{\hat{f}(X_i) - f(X_i)\} \qquad (7.15)\\
&\quad + O_p\left(c_n^2 + h^{2(p+1)}\right),
\end{aligned}
$$

uniformly for d in any compact subset of $\mathbb{R}$. Let $W_i^*(x)$ denote $W_i(x)$ where $\tilde{f}(X_i; x)$ is replaced by $f(X_i)$, i.e., it is the first element of

$$
H_n(x)^{-1}\left[\begin{array}{c} Z_{p,i}(x)\ell_{10}(Y_i, f(X_i), \tau) \\ \ell_{01}(Y_i, f(X_i), \tau) \end{array}\right].
$$

We point out here that $E[W_i^*(x)|X_i = x] = 0$. Since $\tilde{f}(X_i; x) - f(X_i) = O\{(X_i - x)^{p+1}\}$ by the definition of $\tilde{f}(u; x)$, (7.14) implies that the second term of (7.15) equals

$$
\begin{aligned}
& n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell_{11}(Y_i, f(X_i), \tau + d)W_j(X_i)K_h(X_j - X_i) + O_p\left(c_n^2 + h^{2(p+1)}\right)\\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell_{11}(Y_i, f(X_i), \tau + d)W_j^*(X_i)K_h(X_j - X_i) + O_p\left(c_n^2 + h^{p+1}\right)\\
&\equiv T_n(d) + O_p\left(c_n^2 + h^{p+1}\right)
\end{aligned}
$$

uniformly for d in any compact subset of $\mathbb{R}$. Define W_j^0 to be the first element of the vector

$$
H_n(X_j)^{-1}\left[\begin{array}{c} N_0^{(p)}\ell_{10}(Y_j, f(X_j), \tau) \\ N_{00}\ell_{01}(Y_j, f(X_j), \tau) \end{array}\right].
$$

Note that we again have $E(W_j^0|X_j) = 0$. Write $\rho_{11}(x; d) = E[\ell_{11}(Y, f(X), \tau + d)|X = x]$. Recall that with this definition $\rho_{11}(x; 0) = \rho_{11}(x)$. Then, it can be shown that

$$
T_n(d) = n^{-1}\sum_{j=1}^{n}\rho_{11}(X_j; d)g(X_j)W_j^0 + o_p(n^{-1/2})
$$

uniformly for d in any compact subset of $\mathbb{R}$. The above approximation can be obtained by calculating the second moment of the difference. Note that $\rho_{11}(X_j; d)g(X_j)W_j^0$ is nothing else than an approximation of the conditional mean of $\ell_{11}(Y_i, f(X_i), \tau + d)W_j^*(X_i)K_h(X_j - X_i)$ given (X_j, Y_j) where $j \neq i$.

Since $n^{1/2}c_n^2 \to 0$ and $n^{1/2}h^{p+1} \to 0$ by the condition (B9), we obtain from the above approximations and (7.15) that

$$
Q_{3n}(d) = Q_{4n}(d) + n^{-1}\sum_{j=1}^{n}\rho_{11}(X_j; d)g(X_j)W_j^0 + o_p(n^{-1/2})
$$

29

uniformly for d in any compact subset of $\mathbb{R}$. This entails

$$
\begin{aligned}
n^{1/2}\hat{d} &= \left(-n^{-1}\sum_{i=1}^{n}\ell_{02}(Y_i, f(X_i), \tau)\right)^{-1}\left\{n^{1/2}Q_{4n}(0) + n^{-1/2}\sum_{i=1}^{n}\rho_{11}(X_i)g(X_i)W_i^0\right\} + o_p(1) \\
&= \{E\rho_{02}(X)\}^{-1}n^{-1/2}\sum_{i=1}^{n}\left\{\ell_{01}(Y_i, f(X_i), \tau) + \rho_{11}(X_i)g(X_i)W_i^0\right\} + o_p(1) \\
&= \{E\rho_{02}(X)\}^{-1}n^{-1/2}\sum_{i=1}^{n}\ell^*(Y_i, X_i) + o_p(1).
\end{aligned}
$$

The last approximation follows from the fact that $H_n(x) = D(x)g(x) + o(1)$ uniformly for $x \in \mathcal{S}$. Thus, we conclude that $n^{1/2}\hat{d} \xrightarrow{d} \mathcal{N}_1\left[0, \{E\rho_{02}(X)\}^{-2}E\ell^*(Y, X)^2\right]$.

References

[1] Chambers, R.G., Y. Chung, R. Färe (1996), Benefit and distance function, *Journal of Econonomic Theory*, 70, 407–419.

[2] Debreu, G. (1951), The coefficient of resource utilization, *Econometrica*, 19, 273–292.

[3] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), Generalized partially linear single-index models, *Journal of the American Statistical Association*, 92, 477–489

[4] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

[5] Fan, J., Heckman, N. E and Wand, M. P. (1995), Local polynomial kernel regression for generalized linear models and quasi-likelihood functions, *Journal of the American Statistical Association*, 90, 141–150.

[6] Farrell, M.J. (1957), The Measurement of Productive Efficiency, *Journal of Royal Statistical Society*, Series A, 120:3, 253–281.

[7] Kumar, S. and R.R. Russell (2002), Technological Change, Technological Catch-up, and Capital Deepening: Relative Contributions to Growth and Convergence, *American Economic Review*, 92, 3, 527–548.

[8] Kumbhakar, S. C., Park, B. U., Simar, L. and Tsionas, E. G. (2007), Nonparametric stochastic frontiers: a local maximum likelihood approach, *Journal of Econometrics*, 137, 1–27.

[9] Lewbel, A. and O. Linton, (2002), Nonparametric censored and truncated regression, *Econometrica*, 70, 765–779.

[10] Luenberger, D.G. (1994), Optimality and the theory of value, *Journal of Econonomic Theory*, 63, 147–169.

[11] Maddala, G.S. (2001), *Introduction to Econometrics*, 3rd Edition, John Wiley & Sons Ltd., England.

[12] Simar L. and P. Wilson (2007), Estimation and inference in two-stage, semi-parametric models of production processes, *Journal of Econometrics*, vol 136, 1, 31–64.

[13] Tibshirani, R. and Hastie, T. J. (1987), Local likelihood estimation, *Journal of the American Statistical Association*, 82, 559–567.

[14] Zelenyuk, V. and V. Zheka (2006), Corporate governance and firm's efficiency: the case of a transitional country, Ukraine, *Journal of Productivity Analysis*, 25, 143–168.