I N S T I T U T D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0536

NONPARAMETRIC "REGRESSION" WHEN ERRORS ARE CENTRED AT ENDPOINTS

P. HALL and I. VAN KEILEGOM

http://www.stat.ucl.ac.be

NONPARAMETRIC "REGRESSION" WHEN ERRORS ARE CENTRED AT ENDPOINTS

Peter Hall[1,2] and Ingrid Van Keilegom[2]

ABSTRACT. Increasing practical interest has been shown in regression problems where the errors, or disturbances, are centred in a way that reflects particular characteristics of the mechanism that generated the data. In economics this occurs in problems involving data on markets, productivity and auctions, where it can be natural to centre at an endpoint of the error distribution, rather than at the distribution's mean. Often these cases have an extreme-value character, and in that broader context, examples involving meteorological and record-value data have been discussed in the literature. We shall suggest nonparametric methods for estimating regression curves in these settings, showing that they have features that contrast so starkly with those in better-understood problems that they lead to apparent contradictions. For example, merely by centring errors at their endpoints rather than their means the problem can change from one with a familiar nonparametric character, where the optimal convergence rate is slower than $n^{-1/2}$, to one in the super-efficient class, where the optimal rate is faster than $n^{-1/2}$. Moreover, when the errors are centred in a non-standard way there is greater intrinsic interest in estimating characteristics of the error distribution, as well as of the regression mean itself. The paper will also address this aspect of the problem. The new function-estimation methodology can also be viewed as a competitor of techniques such as data envelopment analysis (DEA) and stochastic frontier analysis (SFA), relative to which it has a greater degree of robustness against outliers.

KEYWORDS. Bandwidth, curve estimation, extreme-value theory, jump discontinuity, kernel, local-linear methods, local-polynomial methods, nonparametric regression, smoothing, super efficiency.

SHORT TITLE. Nonregular nonparametric regression.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2] Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

The research of Van Keilegom was supported by IAP research network grant nr. P5/24 of the Belgian government (Belgian Science Policy). The authors like to thank Léopold Simar for helpful discussions and for providing the data.

1. INTRODUCTION

The problem of estimating the endpoint and tail-shape of a distribution has a distinguished history, not least because it provides important examples of non-regular behaviour for various types of inference. See, for example, Harter and Moore (1965) and Smith (1985). The problem also has important practical motivations, arising in part from the prevalence of power-law distributions; see Zipf (1941, 1949). More recently, endpoint and tail-shape problems have been studied in regression settings, for example in econometric models for auctions.

The importance of endpoint-estimation to auction models, and the consequent fact that statistical inference in such models is non-regular, were first noted by Paarsch (1992) and Donald and Paarsch (1993). The endpoint problem arises there because the distribution of bid price generally depends on all the parameters of the model, for instance on parameters that determine the costs of bidders. For particular examples of auction models, see Paarsch (1992) and Donald and Paarsch (2002).

Similar phenomena occur in truncated- or censored-regression models (e.g. Breen, 1996; Long, 1997), market-structure analysis (e.g. Robinson and Chiang, 1996) and inference for production frontiers in econometrics (e.g. Aigner et al., 1977; Park and Simar, 1994). There is a strong association between these fields and those where extreme-value methods are used; for example, the successful bid at an auction is the extremum of all bids.

Although the term "regression" is commonly used in these settings, strictly speaking it is not correct. Since the error distribution is not centred at its expectation then the "regression mean" no longer admits its conventional definition as the average of the response variable given the value of the covariate, or explanatory, variable. This apparently minor distinction can have a major impact, and for example can lead to an intriguing paradox, as we shall show shortly.

In the context of auction models, Hirano and Porter (2003), Jofre-Bonet and Pesendorfer (2003) and Chernozhukov and Hong (2004) studied parametric approaches to inference about distribution endpoints and jump heights. Related statistical work is more in the setting of parametric regression; see, for example, Koenker et al. (1994), Smith (1994), Jurečková (2000), Portnoy and Jurečková (2000) and Knight (2001). However, it is feasible to take a nonparametric view of this problem, permitting a greater degree of flexibility and generality.

The present paper suggests nonparametric methodology, and describes its properties, in the context of inference about endpoint and tail-shape functions in nonparametric regression. In this case the errors, or disturbances, in the nonparametric model are centred at their endpoints, rather than at their means. The endpoints

may be assumed to take a convenient value such as zero. Thus, the problem of estimating the nonparametric-regression mean becomes that of adaptively estimating the centring function. One of the advantages of the methodology to which this approach leads is its high degree of robustness against outliers, relative to competing methods such as data envelopment analysis (DEA) and stochastic frontier analysis (SFA).

Estimation of characteristics of the error distribution is sometimes also of practical interest. This problem can have several forms, depending on the extent of generality required. For example, if the error distribution has a jump discontinuity at its endpoint then the height of the jump can be treated nonparametrically, or modelled parametrically, as a function of the explanatory variable. The endpoint might be approached in a polynomial way, and then the exponent, or degree, may be one of the subjects of inference. This paper will address those issues, too.

The problem of nonparametric regression with endpoint-centred errors also has significant theoretical motivation. In particular, depending on the way in which the endpoint is approached, substantially faster convergence rates can be achieved than in conventional settings. For example, suppose we observe $Y_i = a(X_i) + \epsilon_i$ for $1 \le i \le n$, where the errors ϵ_i are independent and identically distributed with a distribution that has a jump discontinuity at one of its endpoints, and also has finite variance; and a denotes a twice-differentiable function. The estimator of a given in this paper has root-mean-square convergence rate $n^{-2/3}$, which beats even the rate $n^{-1/2}$ for a parametric setting, let alone the rate $n^{-2/5}$ for standard nonparametric regression with twice-differentiable functions. We shall show that the rate $n^{-2/3}$ is minimax-optimal.

However, it is well-known that the rate $n^{-2/5}$ is also minimax-optimal, for estimating the same function. How can this be? This paradox can be resolved by noting that the two functions being estimated are not quite identical. They differ by a constant, equal to the difference, δ say, between the mean and the endpoint of the error distribution. The constant cannot be estimated at a faster rate than $n^{-2/5}$. However, this explanation is not without its own element of surprise, since it might be thought that estimation of δ would be a semiparametric rather than a nonparametric problem; if we could observe the errors directly then we could estimate their endpoint at rate n^{-1} and their mean at rate $n^{-1/2}$, both expressed in root-mean-square terms.

2. METHODOLOGY

2.1. Model. Assume that data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated by the model

$$Y_i = a(X_i) + \epsilon_i \,, \tag{2.1}$$

where a denotes a smooth function, each X_i is a p-vector and each Y_i is a scalar. It is supposed that the distribution of the error, or disturbance, ϵ_i, conditional on $X_i = x$, has density $f(\cdot \mid x)$, with the property that $f(u \mid x) = 0$ for $u < 0$ and

$$f(u \mid x) = b(x)\, c(x)\, u^{c(x)-1} + O\bigl(u^{c(x)+d-1}\bigr) \quad \text{as} \quad u \downarrow 0\,, \tag{2.2}$$

where $0 < d < \infty$. The quantities b and c are smooth, strictly positive functions from $\mathbb{R}^p$ to $\mathbb{R}$. We wish to estimate a, and sometimes also b and c.

2.2. Nonparametric estimation of a. Let $h > 0$ denote a bandwidth. Given $x \in \mathbb{R}^p$, let $\mathcal{S}(x, h)$ be the set of pairs (α, β), where α is a scalar and β is a p-vector, such that $Y_i \geq \alpha + \beta^{\mathrm{T}}(X_i - x)$ for all indices i with $\|X_i - x\| \leq h$. Our initial estimator of $a(x)$ is

$$\tilde{a}(x) = \sup\{\alpha : (\alpha, \beta) \in \mathcal{S}(x, h)\}\,. \tag{2.3}$$

The one-sided nature of inference in this problem raises interesting issues connected with existence of the estimator, and edge effects. To appreciate why, consider the case where the points X_i, for $1 \leq i \leq n$, all lie in a p-variate half-space defined by an infinite plane passing through x. Then, there exists β such that $\beta^{\mathrm{T}}(X_i - x) < 0$ for $1 \leq i \leq n$. Since the length of β can be chosen arbitrarily large without altering the sign property, then $\tilde{a}(x)$, defined at (2.3), equals $+\infty$.

Let $\mathcal{R}$ denote the support of the common density, g_X say, of the X_i's, and write $\partial\mathcal{R}$ for the boundary of $\mathcal{R}$. If g_X is continuous and positive in $\mathcal{R}$, and if x is distant at least sh, where $s > 0$, from $\partial\mathcal{R}$, then the probability that $\tilde{a}(x) = +\infty$ converges to zero exponentially fast, as a function of n, as the latter increases. See section 5.1. However, if x lies exactly on $\partial\mathcal{R}$, then, depending on the shape of the boundary, the probability can equal 1, even for finite n. Details are given in section 3.1.

Arguably the simplest way of overcoming these difficulties is to set an upper bound, B say, on the largest value that $a(x)$ can take, and estimate $a(x)$ by averaging $\tilde{a}(u)$ over all values of u for which $|x - u| \leq h_1$ and $|\tilde{a}(u)| \leq B$, where h_1 is another bandwidth. We shall discuss this approach in the next paragraph. Another method, more difficult to implement, is to distort the region of radius h centred at x, within which X_i must lie in order for (X_i, Y_i) to be used to construct $\tilde{a}(x)$, so that the region includes values of X_i that are further than h from x and appropriately complement the values of X_i that are within h of x.

One form that the averaging of $\tilde{a}(u)$ can take is based on local-linear smoothing. There we choose $\hat{\alpha}_1 = \alpha_1 \in \mathbb{R}$ and $\beta_1 \in \mathbb{R}^p$ to minimise

$$\int \bigl\{\tilde{a}(x + h_1 u) - \alpha_1 - \beta_1^{\mathrm{T}} u\bigr\}^2\, I\{|\tilde{a}(x + h_1 u)| \leq B\}\, K(u)\, du\,, \tag{2.4}$$

where K is a bounded, spherically-symmetric probability density supported on the p-variate unit sphere centred at the origin, and h_1 is another bandwidth. Then we put $\hat{a}(x) = \hat{\alpha}_1$.

Alternatively, we may define

$$\check{a}(x) = \frac{\int_{\mathcal{R}(x)} \tilde{a}(x + h_1 u)\, I\{|\tilde{a}(x + h_1 u)| \le B\}\, K(u)\, du}{\int_{\mathcal{R}(x)} I\{|\tilde{a}(x + h_1 u)| \le B\}\, K(u)\, du}\,, \tag{2.5}$$

where $\mathcal{R}(x)$ denotes the set of points $u \in \mathbb{R}^p$ such that $x + h_1 u \in \mathcal{R}$. Both these approaches also overcome problems caused by discontinuities in the function $\tilde{a}$. While both address the issue of boundary effects, the estimator $\hat{a}$ suffers less from boundary bias than $\check{a}$. In both $\hat{a}$ and $\check{a}$ we may use a soft-thresholding approach to inclusion of values of u for which $|\tilde{a}(x + u)| \le B$, rather than the hard thresholding suggested by (2.4) and (2.5).

Plug-in methods can be used to choose the bandwidth, h, empirically. However, motivation for that technique requires theory about large-sample properties of $\tilde{a}$, and so discussion of empirical bandwidth selection is deferred to sections 2.5 and 3.2.

General polynomial-optimisation methods can be employed to estimate a, although at the expense of substantially greater computational labour. Given a degree $q \ge 1$ for the polynomial, we might take $\mathcal{S}(x, h)$ to be the class of all scalar parameters α and $\beta_r(j_1, \ldots, j_r)$, for $1 \le r \le q$ and $1 \le j_1, \ldots, j_r \le p$, such that

$$Y_i \ge \alpha + \sum_{r=1}^{q} \frac{1}{q!} \sum_{1 \le j_1, \ldots, j_r \le p} \beta_r(j_1, \ldots, j_r)\, (X_i - x)_{j_1} \ldots (X_i - x)_{j_r} \tag{2.6}$$

for all i with $\|X_i - x\| \le h$. Taking β to denote the vector of all components $\beta_r(j_1, \ldots, j_r)$ for $1 \le r \le q$ and $1 \le j_1, \ldots, j_r \le p$, we again define $\tilde{a}(x)$ by (2.3), define $\hat{a}(x) = \alpha_1$ by minimising (2.4), and define $\check{a}(x)$ by (2.5).

The local-linear estimator introduced in the first paragraph of this section can be viewed as based on a local, functional version of a linear-programming algorithm. See Smith (1994) and Portnoy and Jurečková (2000) for related methodologies. The more general estimator, introduced in the paragraph above, requires polynomial programming for implementation.

2.3. Nonparametric estimation of b and c. In principle, completely nonparametric methods may be used to estimate the functions b and c, although in practice one would often take c to be a constant, rather than a nondegenerate function of x.

When estimating b and c we need not use the numerical value of h employed for $\tilde{a}$. However, in the brief account below we shall continue to use the notation h. Define the residuals $\tilde{\epsilon}_i = Y_i - \tilde{a}(X_i)$, and let $\mathcal{T}(x, h)$ denote the set of $\tilde{\epsilon}_i$'s for which

$\tilde{\epsilon}_i > 0$ and $\|X_i - x\| \le h$. Put $N_1 = \#\mathcal{T}(x,h)$, and rank the elements of $\mathcal{T}(x,h)$ as $0 < \hat{\epsilon}_{(1)}(x,h) \le \ldots \le \hat{\epsilon}_{(N_1)}(x,h)$. Put

$$\hat{c}(x) = \left\{ \log \hat{\epsilon}_{(r+1)}(x,h) - \frac{1}{r} \sum_{i=1}^{r} \log \hat{\epsilon}_{(i)}(x,h) \right\}^{-1}, \; \hat{b}(x) = (r/N_1) \left\{ \hat{\epsilon}_{(r+1)}(x,h) \right\}^{-\hat{c}(x)},$$

where r, another smoothing parameter, denotes a threshold.

Optimal choice of bandwidth for estimating b and c is a highly complex matter. While it depends to some extent on the level of smoothing used to compute residuals, it is also influenced by issues, such as the value of d in (2.2), which have no direct bearing on choice of h when smoothing to estimate a. Therefore an account of optimal bandwidth selection when estimating b and c has a number cases, determined by the intersection of circumstances for b and c with those for a. We shall not discuss the matter further here.

The estimators $\hat{b}$ and $\hat{c}$ above can be thought of as local, function versions of conditional maximum-likelihood estimators suggested by Hill (1975).

2.4. Alternative approaches to estimating b and c. Here we suggest quasi-parametric estimators of b and c, starting from the nonparametric estimators of a suggested in section 2.2.

If the sample size were ν; if we were to observe all ν values of ϵ_i which did not exceed a threshold, t say; and if there were just r of these, denoted by $\epsilon_{(1)} < \ldots < \epsilon_{(r)}$, with $\epsilon_{(i)}$ having density g_i and distribution function G_i; then the conditional likelihood of these data would be proportional to

$$\left\{ \prod_{i=1}^{r} g_i(\epsilon_{(i)}) \right\} \prod_{i=r+1}^{\nu} \{1 - G_i(t)\} .$$

Writing $X_{(i)}$ for the concomitant of $\epsilon_{(i)}$; assuming, in reflection of property (2.2), that $g_i(u) = b(X_{(i)} \mid \theta)\, c\, u^{c-1}$ for $0 < u \le t$, where $b(\,\cdot \mid \theta)$ is a model for the scaling "function of proportionality," b, determined by a finite parameter vector, θ; and, for simplicity, taking c to be a scalar rather than a function; we may estimate c and θ by maximising

$$\left\{ \prod_{i=1}^{r} b(X_{(i)} \mid \theta)\, c\, \epsilon_{(i)}^{c-1} \right\} \prod_{i=r+1}^{\nu} \left\{ 1 - b(X_{(i)} \mid \theta)\, t^c \right\} .$$

Differentiating with respect to c and θ gives, respectively, the equations

$$(\log t) \sum_{i=r+1}^{\nu} \frac{t^c\, b(X_{(i)} \mid \theta)}{1 - t^c\, b(X_{(i)} \mid \theta)} - \sum_{i=1}^{r} \log \epsilon_{(i)} = r/c , \tag{2.7}$$

$$\sum_{i=1}^{r} \frac{\partial b(X_{(i)} \mid \theta)/\partial\theta}{b(X_{(i)} \mid \theta)} - \sum_{i=r+1}^{\nu} \frac{t^c\, \partial b(X_{(i)} \mid \theta)/\partial\theta}{1 - t^c\, b(X_{(i)} \mid \theta)} = 0 . \tag{2.8}$$

Of course, in practice we do not observe the ϵ_i's. We replace them by the values of strictly positive residuals $\tilde{\epsilon}_i$, defined as in section 2.3 and ranked as $0 < \tilde{\epsilon}_{(1)} \le \tilde{\epsilon}_{(2)} \le \ldots$. (Recall that the ranking $\hat{\epsilon}_{(1)}(x,h) \le \hat{\epsilon}_{(2)}(x,h) \le \ldots$, introduced in section 2.3, is only a local ranking of positive residuals.) For the threshold we take $t = \tilde{\epsilon}_{(r+1)}$, i.e. the smallest strictly positive $\tilde{\epsilon}_i$ not previously considered. There are thus only two smoothing parameters, r and the value of h used to estimate a prior to computing the residuals.

In some applications the value of c would be known. For example, a parametric model, the detailed constraints of which we might wish to avoid, could assert that the error distribution has a jump discontinuity at its endpoint. For example, this is the case for the auction models discussed in section 1. In such instances, $c = 1$. More generally, if c were known we would use only equations (2.8), and choose $b(x \mid \theta)$ to be a relatively simple model, for example log-linear in x.

In principle, a plug-in rule could be developed for choosing the smoothing parameters h and r in this setting. However, several subsidiary smoothing parameters would be needed in order to select the two main ones. A more attractive proposition is to experiment with values that would be appropriate if the data were generated by a relatively simple model, for example having a log-linear function b and gamma or Weibull distributions for the errors.

2.5. Outline of theoretical properties. We shall show in section 3 that, when constructing the local-linear estimator $\tilde{a}$ and its smoothed versions $\hat{a}$ and $\check{a}$, it is generally optimal to choose $h \sim \text{const.}\, n^{-1/(p+2c)}$. In this case the estimators have root-mean-square convergence rate $n^{-2/(p+2c)}$, when applied to cases where a has two derivatives. For very general choices of the error distribution, this rate is optimal when $0 < c < 2$. Even if the functions b and $c \in (0,2)$ take known, constant values, and we know the error distribution exactly, for example that it is gamma or Weibull, the rate $n^{-2/(p+2c)}$ cannot be improved upon.

However, when $c \ge 2$, and we have sufficient information about the error distribution, the convergence rate of estimators of a can be improved by using other approaches. For instance, if b and c are constant, and if the error density f is known, then an estimator of a that is based on maximising a "local" version of log-likelihood can produce an estimator that converges to a at rate $n^{-2/(p+4)}$, rather than $n^{-2/(p+2c)}$, when $p > 2$ and a has two derivatives.

The problem is more awkward when the error distribution is not known. There, the convergence rate $n^{-2/(p+2c)}$ can be close to optimal. In particular, if we know only that the errors have a common density f, with $f(u) = b\,c\,u^{c-1} + O(u^{c+d})$ as $u \downarrow 0$, where $b, c > 0$ are fixed constants, then the minimax-optimal convergence

rate of estimators of a is $n^{-2/(p+2c)-\delta(d)}$, where $\delta(d) > 0$ converges to zero as $d \downarrow 0$.

Moreover, the approach to regression that centres the residuals at an endpoint of the support, rather than at the mean, is more difficult to motivate when c is relatively large. If the error density decreases to zero very smoothly and gradually in its tails, then knowing one of the ends of its support is less important than it would be if the tail ended relatively abruptly. This consideration, and that in the previous paragraph, suggest that it is reasonable to confine attention to the estimators of a considered in section 2.2.

The convergence rate of $n^{-2/(p+4)}$, mentioned in connection with the case $c > 2$ and a known error distribution, is optimal when estimating p-variate regression functions that are twice-differentiable, provided the error distribution is centred at its mean rather than its endpoint. See e.g. Stone (1980, 1982). The rate of $n^{-2/(p+2c)}$ that we obtain when $0 < c < 2$ is of course faster. This improvement in the rate of convergence is another reason for paying special attention to the case $0 < c < 2$.

3. THEORETICAL PROPERTIES

3.1. Convergence rates of estimators of a. Assume that data (X_i, Y_i) are generated by the model at (2.1), where

> the p-variate explanatory variables $\mathcal{X} = \{X_1, X_2, \ldots\}$ are independent; conditional on the X_i's, the errors $\epsilon_1, \epsilon_2, \ldots$ are independent, and the marginal density f_i of ϵ_i depends on $\mathcal{X}$ only through X_i; the X_i's are identically distributed as X, the density of which is supported in a compact region, $\mathcal{R} \subseteq \mathbb{R}^p$, and is continuous and nonzero there; $f_i = f(\cdot \mid X_i)$, where f satisfies (2.2), $d > 0$ is fixed, b and c are Hölder-continuous functions satisfying $C_1 \le b(x), c(x) \le C_2$ for all $x \in \mathcal{R}$, C_1 and C_2 are constants satisfying $0 < C_1 < C_2 < \infty$, and the remainder in (2.2) is of that order uniformly in $x \in \mathcal{R}$; and $\sup_{i,x} E(|\epsilon_i|^{2+\eta} \mid X_i = x) < \infty$ for some $\eta > 0$. (3.1)

Recall from section 2 that the one-sided nature of the inference problem means that the estimator $\tilde{a}$ will often tend not to be defined at the boundary. However, $\tilde{a}$ may be well-defined very close to the boundary. To elucidate this behaviour we shall consider two types of x:

> Either x is fixed, as an interior point of $\mathcal{R}$, or x is close to the boundary, $\partial\mathcal{R}$, of $\mathcal{R}$, and varying with n. In the latter case we ask that $x = x(n) = x_0 + v(x_0)\, sh$, where $x_0 \in \partial\mathcal{R}$, $s > 0$, and $v(x_0)$ is the normal to the tangent plane to $\partial\mathcal{R}$ at x_0, oriented so that it points into $\mathcal{R}$. In this case, x_0 and s are held fixed. (3.2)

If $x \in \mathcal{R}$ is an interior point, or if $x = x(n) = x_0 + v(x_0)\, sh$ where $x_0 \in \partial\mathcal{R}$ and $s \ge 1$, let $\mathcal{U}(x)$ denote the closed, p-variate sphere of unit radius centred at

the origin. If $x = x(n) = x_0 + v(x_0)\, sh$ with $x_0 \in \partial\mathcal{R}$ and $0 < s < 1$, take $\mathcal{U}(x)$ to be the larger of the two parts of the just-mentioned sphere that are obtained by cutting it by the plane that is perpendicularly distant s from the origin and has its normal in the direction $v(x)$, pointing towards the centre to the sphere.

Let $\dot{a}$ and $\ddot{a}$ denote the p-vector of first derivatives, and $p \times p$ matrix of second derivatives, of the function a, and suppose that

$$\text{the function } a \text{ has two continuous derivatives in } \mathcal{R}, \text{ and if } x = x_0 + v(x_0)\, sh \text{ then } \partial\mathcal{R} \text{ has a continuously turning tangent plane at } x_0. \tag{3.3}$$

Assume too that

$$\text{for some } 0 < \eta < 1/(2p) \text{ and all sufficiently large } n,\ n^{\eta-(1/p)} < h < n^{-\eta}. \tag{3.4}$$

Given $x \in \mathcal{R}$, let $E_1, E_2, \ldots$ denote independent, exponentially distributed random variables, all with unit mean, write γ for Euler's constant, and define

$$Z_j(x) = \exp\left[- c(x)^{-1} \left\{ \sum_{i=j}^{\infty} (E_i - 1)\, i^{-1} + \gamma - \sum_{i=1}^{j-1} i^{-1} \right\} \right], \quad j \ge 1\,. \tag{3.5}$$

Given $x \in \mathcal{R}$, let $U_1(x), U_2(x), \ldots$ be independent and identically distributed random p-vectors, independent too of the $Z_j(x)$'s, and uniformly distributed on $\mathcal{U}(x)$. For $c_1, c_2 \ge 0$, define

$$Q_1(c_1, c_2 \mid x) = \sup_{\beta \in \mathbb{R}^p} \inf_{1 \le i < \infty} \Big[c_1 \left\{ \beta^{\mathrm{T}}\, U_i(x) + \tfrac{1}{2}\, U_i(x)^{\mathrm{T}}\, \ddot{a}(x)\, U_i(x) \right\} + c_2\, b(x)^{-1/c(x)}\, Z_i(x) \Big]\,.$$

In the statement of Theorem 1 below, we let x_1 denote x if x is an interior point of $\mathcal{R}$, and $x_1 = x_0$ if $x = x(n) = x_0 + v(x_0)\, sh$. Let $w(p)$ be the content of the p-variate unit sphere (thus, $w(1) = 2$, $w(2) = \pi$), let $g_X(x)$ represent the value of the density of the distribution of X at x, and put $w_x = w(p)\, g_X(x_1)$. (To simplify notation we suppress the role of x_1 here.) We use a simpler rule than that in section 2.2 to take care of cases where $\tilde{a}(x)$ is infinite. However, the last sentence in the theorem remains true if we define $\tilde{a}(x)$ to equal zero whenever $|\tilde{a}(x)| > B$, provided $B > |a(x)|$.

Theorem 1. *Assume* (3.1)–(3.4). *(a) If* $(w_x n h^p)^{1/c(x_1)}\, h^2 \to \rho$, *where* $\rho \in [0, \infty)$, *then* $(w_x n h^p)^{1/c(x_1)}\, \{\tilde{a}(x) - a(x)\} \to Q_1(\rho, 1 \mid x_1)$ *in distribution. (b) If* $(n h^p)^{1/c(x_1)} \times h^2 \to \infty$ *then* $h^{-2}\, \{\tilde{a}(x) - a(x)\} \to Q_1(1, 0 \mid x_1)$ *in distribution. Furthermore, if we take the precaution of defining* $\tilde{a}(x)$ *to equal an arbitrary but fixed constant in*

cases where it would otherwise be infinite, then second moments converge to those of the limiting distributions.

Proofs of Theorems 1–3 will be given in section 5. It is crucial, in condition (3.2), that we take $s > 0$ rather than $s \geq 0$. If $s = 0$ then x lies right on the boundary of $\mathcal{R}$, and in such cases the theorem is false. For example, if $\mathcal{R}$ is a convex region with a smooth boundary, such as a sphere, then with probability 1, $\tilde{a}(x_0) = \infty$ for all $x_0 \in \partial\mathcal{R}$. However, it follows from the theorem that for points x that are arbitrarily close to $\partial\mathcal{R}$, on the scale of the bandwidth, without being right on the boundary, the probability that $\tilde{a}(x)$ is finite converges to 1, and in fact the estimator $\tilde{a}(x)$ attains optimal convergence rates.

Asymptotic properties of $\hat{a}$ and $\check{a}$ are similar, except that the limiting distribution of $\hat{a}$ is more tedious to define. Therefore we shall confine ourselves to $\check{a}$. To further abbreviate our treatment we shall restrict attention to the case where

$$\begin{aligned} &x \text{ is an interior point of } \mathcal{R},\ h_1 = th \text{ for a fixed constant } t > 0, \text{ and} \\ &(w_x n h^p)^{1/c(x)}\, h^2 \to \rho \in [0, \infty). \end{aligned} \tag{3.6}$$

Let $Z_1, Z_2, \ldots$ be as at (3.5); for simplicity we drop the argument x. Re-define $U_1, U_2, \ldots$ to be independent, of one another and of the Z_j's, and uniformly distributed in the p-variate sphere of radius $t + 1$ centred at the origin. Given a p-vector u with $\|u\| \leq t$, let $(S_1(u), T_1(u)), (S_2(u), T_2(u)), \ldots$ denote the values $(U_{i_1(u)}, Z_{i_1(u)}), (U_{i_2(u)}, Z_{i_2(u)}), \ldots$ of $(U_i, Z_i) = (U_i, Z_i(x))$ for which $\|U_i - u\| \leq h$, ordered such that $Z_{i_1(u)} < Z_{i_2(u)} < \ldots$. With $\kappa = p^{-1} \int \|u\|^2 K(u)\, du$, ∇^2 denoting the Laplacian operator, and $\rho \geq 0$ as in (3.6), define

$$\begin{aligned} Q_2(u \mid x) = \sup_{\beta \in \mathbb{R}^p} \inf_{1 \leq i < \infty} \Big[&\rho \left\{ \beta^{\mathrm{T}} S_i(u) + \tfrac{1}{2}\, S_i(u)^{\mathrm{T}}\, \ddot{a}(x)\, S_i(u) \right\} \\ &+ \left\{ (t+1)^p\, b(x) \right\}^{-1/c(x)} T_i(u) \Big], \\ Q_3(x) = \tfrac{1}{2}\, \rho\, t^2\, \kappa \left(\nabla^2 a \right)(x) + &\int Q_2(u \mid x)\, K(u)\, du\,. \end{aligned}$$

Under conditions (3.1)–(3.6), and taking $B > |a(x)|$ in (2.5), it can be shown that with probability $1 - O(n^{-C})$ for all $C > 0$, the estimator $\check{a}(x)$, at (2.5), satisfies

$$\check{a}(x) = \int \tilde{a}(x + h_1 u)\, K(u)\, du\,. \tag{3.7}$$

Theorem 2 applies with equal validity to the estimators at (2.5) and (3.7).

Theorem 2. *Assume* (3.1)–(3.6)*, and that the kernel K, used to define $\check{a}(x)$, is a bounded, spherically-symmetric probability density supported on the unit sphere centred at the origin. Then $(w_x n h^p)^{1/c(x)} \{\check{a}(x) - a(x)\} \to Q_3(x)$ in distribution.*

Furthermore, if in the integrand at (3.7) *we take the precaution of defining* $\check{a}(x+h_1u)$ *to equal an arbitrary but fixed constant in cases where it would otherwise be infinite, then the second moment converges to that of the limiting distribution.*

3.2. Choice of bandwidth. Theorems 1 and 2 imply that, except in pathological cases where $\ddot{a}(x) = 0$, the optimal convergence rate of $\tilde{a}(x)$ and $\check{a}(x)$ to $a(x)$ is achieved by choosing the bandwidth, h, so that $(nh^p)^{-1/c(x)}$ and h^2 are of the same size, and in particular, $h \sim \text{const.}\, n^{-1/\{p+2c(x)\}}$. If x does not lie on the boundary of $\mathcal{R}$, and if $(w_x nh^p)^{1/c(x)}\, h^2 \to \rho \in [0,\infty)$, then the asymptotic mean squared error of $\tilde{a}(x)$ is given by

$$Q_4(\rho \mid x) = E\biggl\{ \sup_{\beta\in\mathbb{R}^p} \inf_{1\le i<\infty} \Bigl[\rho\left\{\beta^{\mathrm{T}} U_i + \tfrac{1}{2}\, U_i^{\mathrm{T}}\, \ddot{a}(x)\, U_i\right\} + b(x)^{-1/c(x)}\, Z_i(x)\Bigr]\biggr\}^2 , \tag{3.8}$$

where $U_1, U_2, \ldots$ are uniformly distributed on the unit sphere centred at the origin, $Z_1(x), Z_2(x), \ldots$ are defined at (3.5), and the U_i's and $Z_i(x)$'s are completely independent. Therefore, if $h = w_x^{-1/\{p+2c(x)\}}\, \rho^{1/\{2+p/c(x)\}}\, n^{-1/\{p+2c(x)\}}$ then ρ should ideally be chosen as

$$\rho_0(x) = \operatorname*{argmin}_{\rho}\, Q_4(\rho \mid x)\,. \tag{3.9}$$

In principle, $\ddot{a}(x)$ can be estimated by fitting the polynomial smoother at (2.6) with $q = 2$, and taking the value of $\beta_2(j_1, j_2)$ that results to be our estimator of

$$\ddot{a}_{j_1,j_2}(x) = \frac{\partial^2 a(x)}{\partial x_{j_1} \partial x_{j_2}}\,.$$

However, this approach is highly computer-intensive. A simpler method is to twice numerically differentiate a heavily smoothed version of $\tilde{a}$. Simpler still, if we may make the assumption (A), say, that, for each i, the distribution of ϵ_i does not depend on X_i, then a traditional smoother passed through the data (X_i, Y_i) estimates the value of $\mu(x) = E(Y \mid X = x)$. Under (A), this quantity differs from a only by a constant, and so $\ddot{a} = \ddot{\mu}$. The latter function can be estimated using conventional cubic smoothing. This approach is attractive even if the distribution of ϵ_i depends to some extent on X_i, since it gives a working empirical approximation to $\ddot{a}$.

Methods for estimating $b(x)$ and $c(x)$ were discussed in section 2.2. Substituting these estimators for the true values of $\ddot{a}(x)$, $b(x)$, $c(x)$ and $\rho(h,x)$ in (3.8), we may compute an estimator $\widehat{Q}_4(\rho \mid x)$ of $Q_4(\rho \mid x)$ using Monte Carlo simulation, leading to an estimator $\hat{\rho}_0(x)$ of $\rho_0(x)$ at (3.9). The density of X at x, i.e. $g_X(x)$, can be estimated more conventionally, and thus an estimator $\widehat{w}_x$ of $w_x = w(p)\, g_X(x)$ can be constructed. An empirical bandwidth selector is then given by

$$h(x) = \widehat{w}_x^{-1/\{p+2\hat{c}(x)\}}\, \hat{\rho}_0(x)^{1/\{2+p/\hat{c}(x)\}}\, n^{-1/\{p+2\hat{c}(x)\}}\,.$$

In many circumstances it is feasible to take $c(x)$ to be a constant, not depending on x. Section 2.4 discussed inference in this setting. Then a global approach to bandwidth choice is possible, as follows. We shall proceed as though the density g_X is constant; if it is not, using its average value, rather than attempting to accommodate its variation, greatly simplifies matters. Thus, we take $\widehat{w}$ to be an estimator of the average value of w_x. The mean integrated squared error of $\tilde{a}(x)$ is asymptotic to $Q_4(\rho) = \int_{\mathcal{R}} Q_4(\rho \mid x)\, dx$, of which an estimator is $\widehat{Q}_4(\rho) = \int_{\mathcal{R}} \widehat{Q}_4(\rho \mid x)\, dx$, leading to an estimator $\hat{\rho}_0 = \operatorname{argmin}_\rho \widehat{Q}_4(\rho)$ of $\rho_0 = \operatorname{argmin}_\rho Q_4(\rho)$. A global bandwidth for constructing $\tilde{a}$ is thus $h = \widehat{w}^{-1/(p+2\hat{c})}\, \hat{\rho}_0^{1/(2+p/\hat{c})}\, n^{-1/(p+2\hat{c})}$.

3.3. Optimality. We shall show in this section that the convergence rates implied by Theorems 1 and 2, and also lower bounds of the same orders, are available uniformly over classes $\mathcal{A}$ of functions a with two bounded derivatives. The possibility that either the proportionality constant, b, or the exponent, c, varies with the design variable, X_i, is not relevant to discussion of the lower bound, and for this reason, for simplicity, and since our lower-bound results are stronger if we narrow the class of error distributions for which worst-case performance is achieved, we shall take the distribution of $\epsilon = \epsilon_i$ to be a single, specific one, say the gamma:

$$f(u) = f_i(u) = \frac{1}{\Gamma(c)}\, u^{c-1}\, e^{-u}\,, \quad \text{where } c > 0 \text{ is fixed}\,. \tag{3.10}$$

In the lower-bound calculations, $c > 0$ will be assumed known.

Likewise, we shall treat just one distribution of $X = X_i$ and one region $\mathcal{R}$. In particular, writing $\mathcal{V}(x, r)$ for the closed sphere centred at x and of radius $r > 0$, we shall assume that

$$\mathcal{R} = \mathcal{V}(x_0, 1) \text{ and } X \text{ is uniformly distributed on } \mathcal{R}. \tag{3.11}$$

Given $C > 0$, let $\mathcal{A} = \mathcal{A}(C)$ denote the class of functions a for which first and second derivatives exist and are bounded absolutely by C, let $\bar{\mathcal{A}}$ denote the class of bounded functions $\bar{a}$ of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, the latter generated as at (2.1), and let $\mathcal{R}_h$ be the set of all points in $\mathcal{R}$ that are distant at least h from $\partial\mathcal{R}$.

Theorem 3. *Assume* (3.10) *and* (3.11)*, and, when constructing* $\tilde{a}(x)$*, let* $h = \text{const.}\, n^{-1/(p+2c)}$*, except that we take* $\tilde{a}(x)$ *equal to an arbitrary but fixed constant in cases where it would otherwise be infinite. Then,*

$$\sup_{x \in \mathcal{R}_h} \sup_{a \in \mathcal{A}} E\{\tilde{a}(x) - a(x)\}^2 = O\big(n^{-2/(p+2c)}\big) \tag{3.12}$$

as $n \to \infty$*. Furthermore, if* $0 < c < 2$*,*

$$\liminf_{n \to \infty} n^{2/(p+2c)} \inf_{\bar{a} \in \bar{\mathcal{A}}} \sup_{a \in \mathcal{A}} E\{\bar{a}(x) - a(x)\}^2 > 0 \quad \textit{for each } x \in \mathcal{R} \setminus \partial\mathcal{R}\,, \tag{3.13}$$

$$\liminf_{n\to\infty} n^{2/(p+2c)} \inf_{\bar{a}\in\bar{\mathcal{A}}} \sup_{a\in\mathcal{A}} \int_{\mathcal{R}_h} E\{\bar{a}(x) - a(x)\}^2\, dx > 0\,. \tag{3.14}$$

Together, (3.12)–(3.14) imply that the estimator $\tilde{a}$ achieves the minimax-optimal rate, $n^{-2/(p+2)}$, uniformly over all functions $a \in \mathcal{A}$, and that the optimality can be expressed in either a local or a global sense. Similarly, it may be proved that if $\mathcal{A}_q$ is taken to be the class of functions a with $q+1$, rather than 2, bounded derivatives, then the qth degree local-polynomial approach, discussed in section 2.2, achieves the minimax-optimal convergence rate of $n^{-2q/(p+2cq)}$), uniformly over functions in $\mathcal{A}_q$.

4. NUMERICAL PROPERTIES

4.1. Simulations. Consider independent and identically distributed data (X_i, Y_i) $(1 \le i \le n)$ satisfying the model $Y_i = a(X_i) + \epsilon_i$ given in (2.1). The covariate X_i has a uniform distribution on the interval $[0, 1]$. We consider three models for $a(x)$ $(0 \le x \le 1)$:

$$\begin{aligned} &\text{Model 1}: a(x) = 10\,(x - a_0)^3, && a_0 = 0.25, 0.5 \\ &\text{Model 2}: a(x) = \exp(-a_0 x^2), && a_0 = 1, 2 \\ &\text{Model 3}: a(x) = a_0\,\cos(\pi x), && a_0 = 0.25, 0.5. \end{aligned} \tag{4.1}$$

Figure 1 shows the graphs of these six frontier functions. The error ϵ_i is taken from a Gamma distribution:

$$f(u|x) = \frac{1}{s(x)^c\,\Gamma(c)}\; u^{c-1}\,\exp\{-u/s(x)\}$$

$(u \ge 0)$, where $s(x) = 1 + 2x$ and $c = 0.5, 1$ or 1.5. These three values of c are such that, as $u \downarrow 0$, $f(u|x) \to \infty$, $f(u|x) \to s(x)^{-1}$ and $f(u|x) \to 0$ respectively. Note that this density is of the general type (2.2), with $b(x) = \{c\,s(x)^c\,\Gamma(c)\}^{-1}$.

The simulations are executed based on 100 arbitrary samples of size $n = 200$ and $n = 400$. For each sample we estimate $a(x)$ at $x = 0.5$, and use local-linear smoothing to obtain both $\tilde{a}(x)$ and $\hat{a}(x)$. The bandwidth h is calculated from formula (3.8), and we have taken $h_1 = h$. To estimate $\ddot{a}(x)$ we work, as explained in section 3.2, under the working model that the distribution of ϵ_i does not depend on X_i, in which case $\ddot{a}(x)$ equals the second derivative of the regression function $E(Y \mid X = x)$. This second derivative is estimated using local-cubic smoothing, with bandwidth 0.25. The functions $b(x)$ and $c(x) \equiv c$ are estimated employing the procedure explained in section 2.3, where r equals the smallest integer larger than $0.90\,N_1$, and the bandwidth for estimating $b(x)$ and $c(x)$ is chosen as 0.25. The kernel used throughout is the biquadratic kernel, $K(u) = (15/16)\,(1 - u^2)^2\; I(|u| \le 1)$.

Tables 1 and 2 show the estimated bias, variance and mean-squared error (MSE) of $\hat{a}(x)$ at $x = 0.5$, for each of the considered models, as well as the average value of the bandwidth h over the 100 simulation runs, obtained using Monte Carlo simulation of formula (3.8). Note that the functions $a(x)$ considered in this simulation study are neither convex nor concave. In fact, our method imposes neither condition, in contradistinction to, for instance, the DEA (data envelopment analysis) estimator, which requires the function $a(x)$ to be convex.

The tables show that the MSE increases when c increases, which is to be expected since the higher the value of c, the smaller the density $f(\cdot \mid x)$ of the error close to the frontier, and so the harder the estimation of the frontier. These findings also agree with the theoretical results of section 3. This sparsity of data close to the frontier affects especially the bias of the estimator, since it is clear that the estimator $\tilde{a}(x)$ tends to overestimate $a(x)$ whenever there are few observations near the boundary. Finally, comparing Tables 1 and 2 we see that both the bias and the variance decrease as the sample size increases.

4.2. Data analysis. We consider data on 123 American electric utility companies, studied by Christensen and Greene (1976), Greene (1990) and Hall and Simar (2002), among others. We focus here on the relation between $Y_i = -\log(C_i/P_i)$ and $X_i = \log(Q_i)$, where C_i is the cost, Q_i the output and P_i the price of fuel for each company. We fit the model

$$Y_i = a(X_i) + \epsilon_i \,,$$

where it is assumed that the conditional density of the errors ϵ_i satisfies relation (2.2). The scatterplot of the data, together with the estimated frontier curve $\hat{a}(x)$, is shown in Figure 2. We restrict the region of estimation to $[4.6, 11.2]$, to avoid estimation in sparse areas of X. Both the estimation of $\tilde{a}(x)$ and $\hat{a}(x)$ is done using local-linear smoothing. At each point of an equispaced grid of 34 values between 4.6 and 11.2 we estimate the bandwidth $h = h_1$ from formula (3.8), yielding values in the range from 0.54 to 1.06. The bandwidth for estimating $\ddot{a}(x)$, $b(x)$ and $c(x)$ is chosen as one fifth of the total range, namely 1.32, whereas to estimate the design density, we use kernel estimation based on the normal reference rule. The kernel used throughout is again the biquadratic kernel.

Figure 2 suggests that a linear model is appropriate for these data. However, it is particularly satisfying to reach that conclusion using a highly adaptive method which does not impose linearity, or even convexity, as a prior assumption.

5. TECHNICAL ARGUMENTS

5.1. Proof of Theorem 1. To simplify notation we shall assume that $w_x = 1$ throughout; this can always be achieved via a change of scale. For brevity we shall deal only with the case where x is an interior point of $\mathcal{R}$. Put $\gamma_\alpha(x) = a(x) - \alpha$, a scalar, and $\gamma_\beta(x) = h^{-1}\{\dot{a}(x) - \beta\}$, a p-vector. Let $\mathcal{I}(x,h)$ denote the set of indices i such that $\|X_i - x\| \le h$, and for $i \in \mathcal{I}(x,h)$, define $V_i = h^{-1}(X_i - x)$. In this notation,

$$Y_i - \alpha - \beta^{\mathrm{T}}(X_i - x) = \gamma_\alpha(x) + h^2\left\{\gamma_\beta(x)^{\mathrm{T}} V_i + \tfrac{1}{2} V_i^{\mathrm{T}} \ddot{a} V_i\right\} + h^2 R_i(x) + \epsilon_i\,,$$

where the remainder, $R_i(x)$, has the property that

$$\sup_{x\in\mathcal{R}} \sup_{i\in\mathcal{I}(x,h)} |R_i(x)| \le R(h) \equiv h^{-2} \sup_{x\in\mathcal{R}} \sup_{u\,:\,\|u\|\le 1,\, x+hu\in\mathcal{R}} \Big| a(x+hu) - a(x) - hu^{\mathrm{T}}\dot{a}(x) - \tfrac{1}{2} h^2 u^{\mathrm{T}}\ddot{a}(x)\, u \Big|\,, \quad (5.1)$$

and $R(h) \to 0$ as $h \to 0$.

In particular, asking that $Y_i \ge \alpha + \beta^{\mathrm{T}}(X_i - x)$ for all indices $i \in \mathcal{I}(x,h)$ is equivalent to insisting that

$$\gamma_\alpha + \inf_{i\in\mathcal{I}(x,h)} \left\{h^2\left(\gamma_\beta^{\mathrm{T}} V_i + \tfrac{1}{2} V_i^{\mathrm{T}} \ddot{a} V_i\right) + h^2 R_i + \epsilon_i\right\} \ge 0\,, \quad (5.2)$$

where we have dropped the argument from $\gamma_\alpha(x)$, $\gamma_\beta(x)$, $\ddot{a}(x)$ and $R_i(x)$. Let $\mathcal{S}_1(x,h)$ denote the set of pairs $(\gamma_\alpha, \gamma_\beta)$ such that (5.2) holds, and let $\tilde{\gamma}_1$ denote the infimum of γ_α over $(\gamma_\alpha, \gamma_\beta) \in \mathcal{S}_1(x,h)$. Then, $\tilde{a}(x) = a(x) - \tilde{\gamma}_1$.

It follows from this result and (5.1) that, defining

$$\tilde{\gamma}_2 = \tilde{\gamma}_2(x) = \sup_{\gamma_\beta} \inf_{i\in\mathcal{I}(x,h)} \left\{h^2\left(\gamma_\beta^{\mathrm{T}} V_i + \tfrac{1}{2} V_i^{\mathrm{T}} \ddot{a} V_i\right) + \epsilon_i\right\}, \quad (5.3)$$

and noting that, for any random variable A, essup A is the infimum of constants C for which $P(A \le C) = 1$, we have:

$$h^{-2}\,\mathrm{essup} \sup_{x\in\mathcal{R}} \left|\tilde{a}(x) - a(x) - \tilde{\gamma}_2\right| \to 0\,. \quad (5.4)$$

Defining $N = N(x,h) = \#\,\mathcal{I}(x,h)$, we may write $\tilde{\gamma}_2$ equivalently as

$$\tilde{\gamma}_2 = (nh^p)^{-1/c(x)} \sup_{\gamma_\beta} \inf_{1\le i\le N} \left\{\rho_1\left(\gamma_\beta^{\mathrm{T}} V_{(i)} + \tfrac{1}{2} V_{(i)}^{\mathrm{T}} \ddot{a} V_{(i)}\right) + \xi_{(i)}\right\},$$

where $\rho_1 = (nh^p)^{1/c(x)}\, h^2$, $\xi_{(1)} < \xi_{(2)} < \ldots$ are the ordered values of $(nh^p)^{1/c(x)}\,\epsilon_i$ for $i \in \mathcal{I}$, and $V_{(1)}, V_{(2)}, \ldots$ denote the concomitant values of $V_1, V_2, \ldots$.

For each $r \geq 1$,

$$\text{the limiting joint distribution of } \xi_{(1)}, \ldots, \xi_{(r)} \text{ and } V_{(1)}, \ldots, V_{(r)} \text{ is that of } b(x)^{-1/c(x)}\,(Z_1, \ldots, Z_r) \text{ and } U_1, \ldots, U_r, \text{ where the sequence } Z_1, Z_2, \ldots \text{ is as defined at (3.5) and, independently of the } Z_j\text{'s, the } U_j\text{'s are uniformly distributed on the unit } p\text{-variate sphere.} \tag{5.5}$$

(See Hall, 1978). It may be deduced from (5.5) that, if $\hat{\imath}$ denotes the supremum over integers i_0 for which

$$\sup_{\gamma_\beta} \inf_{1 \leq i \leq N} \left\{ \rho_1 \left(\gamma_\beta^{\mathrm{T}}\, V_{(i)} + \tfrac{1}{2}\, V_{(i)}^{\mathrm{T}}\, \ddot{a}\, V_{(i)} \right) + \xi_{(i)} \right\} < \sup_{\gamma_\beta} \inf_{1 \leq i \leq i_0} \left\{ \rho_1 \left(\gamma_\beta^{\mathrm{T}}\, V_{(i)} + \tfrac{1}{2}\, V_{(i)}^{\mathrm{T}}\, \ddot{a}\, V_{(i)} \right) + \xi_{(i)} \right\},$$

and if $\rho_1 \to \rho \in (0, \infty)$ as $n \to \infty$, then

$$\lim_{i \to \infty} \liminf_{n \to \infty} P(\hat{\imath} \leq i) = 1\,.$$

Therefore, if $\rho_1 \to \rho \in (0, \infty)$ as $n \to \infty$ then

$$(nh^p)^{1/c(x)}\, \tilde{\gamma}_2 \to \sup_{\beta \in \mathbb{R}^p} \inf_{1 \leq i < \infty} \left[\rho \left\{ \beta^{\mathrm{T}}\, U_i + \tfrac{1}{2}\, U_i^{\mathrm{T}}\, \ddot{a}(x)\, U_i \right\} + b(x)^{-1/c(x)}\, Z_i \right]$$

in distribution. The part of Theorem 1 pertaining to $\rho_1 \to \rho \in (0, \infty)$ follows from this property and (5.4).

If $\rho_1 \to 0$ then, since $\xi_{(1)} \to b(x)^{-1/c(x)} Z_1$ in distribution, we have $(nh^p)^{1/c(x)} \times \tilde{\gamma}_2 \to b(x)^{-1/c(x)} Z_1$ in distribution. And if $\rho_1 \to \infty$ then

$$h^{-2}\, \tilde{\gamma}_2 \to \sup_{\beta \in \mathbb{R}^p} \inf_{1 \leq i < \infty} \left\{ \beta^{\mathrm{T}}\, U_i + \tfrac{1}{2}\, U_i^{\mathrm{T}}\, \ddot{a}(x)\, U_i \right\}$$

in distribution. Parts (a) and (b) of Theorem 1 are consequences of these properties.

To establish convergence of second moments it suffices, in view of (5.4), to prove that for some $\eta_1 > 0$,

$$\text{there exist random variables } A_1 \text{ and } A_2 \text{ such that } A_1 \leq \left(nh^p\right)^{1/c(x)} \tilde{\gamma}_2 \leq A_2 \text{ with probability 1, and } E(|A_j|^{2+\eta_1}) \text{ is uniformly bounded for } j = 1, 2. \tag{5.6}$$

It follows from (5.3) that, with $\|\ddot{a}\|$ denoting the supremum of $|v^{\mathrm{T}} \ddot{a} v|$ over unit vectors v, and provided $\mathcal{I}(x, h)$ is not empty,

$$\tilde{\gamma}_2 \geq \inf_{i \in \mathcal{I}(x,h)} \left(\tfrac{1}{2}\, h^2\, V_i^{\mathrm{T}}\, \ddot{a}\, V_i + \epsilon_i \right) \geq \left\{ \inf_{i \in \mathcal{I}(x,h)} \epsilon_i \right\} - \tfrac{1}{2}\, h^2\, \|\ddot{a}\|\,. \tag{5.7}$$

We may take the lower bound to be simply a fixed constant if $\mathcal{I}(x, h)$ is empty, and so, since the latter event has exponentially small probability, then, in proving

finiteness of second moments, we may ignore that case. Denote by $i_1, \ldots, i_N$ the elements of $\mathcal{I}(x,h)$, and put $c_n = \sup_{u\,:\,\|x-u\|\le h}\, c(u)$. Then, for constants $B_1, B_2 > 0$ with $B_1\, B_2^{c(x)} < 1$, and for each $\eta_2 \in (0,1)$,

$$\begin{aligned} E\biggl[\Bigl\{\inf_{i\in\mathcal{I}(x,h)}\,\epsilon_i\Bigr\}^{2+\eta_1}\biggr] &= (2+\eta_1)\int_0^\infty u^{1+\eta_1}\,E\bigl\{P(\epsilon_{i_1} > u \mid X_{i_1})\ldots \\ &\qquad\qquad \times P(\epsilon_{i_N} > u \mid X_{i_N})\bigr\}\,du \\ &= O\biggl\{\int_0^{B_2} u^{1+\eta_1}\,\bigl(1 - B_1\,u^{c_n}\bigr)^{(1-\eta_2)E(N)}\,du\biggr\} \\ &= O\bigl\{(EN)^{-(2+\eta_1)/c_n}\bigr\} = O\Bigl\{\bigl(nh^p\bigr)^{-(2+\eta_1)/c(x)}\Bigr\}\,. \end{aligned} \tag{5.8}$$

From (5.7) and (5.8) we deduce the part of (5.6) pertaining to A_1. To obtain the part of (5.6) for A_2 we shall, for the sake of brevity, assume that the distribution of ϵ is bounded above. The contrary case may be treated using the assumption, in (3.1), that $\sup_{i,x} E(|\epsilon_i|^{2+\eta} \mid X_i = x) < \infty$.

We may write $\epsilon_i = F_i^{-1}(D_i)$, where F_i is the distribution function of ϵ_i conditional on X_i, and $D_1, D_2, \ldots$ are independent and uniformly distributed on the unit interval, independent too of $V_1, V_2, \ldots$. In view of (2.2), there exists $B_3 > 0$ such that $F_i^{-1}(u) \le B_3\,u^{1/c_n}$, uniformly in X_i for which $\|X_i - x\| \le h$. Therefore, by (5.3),

$$\tilde\gamma_2 \le \sup_{\gamma_\beta}\,\inf_{i\in\mathcal{I}(x,h)}\,\Bigl\{h^2\,\bigl(\gamma_\beta^{\mathrm{T}}\,V_i + \tfrac{1}{2}\,V_i^{\mathrm{T}}\,\ddot a\,V_i\bigr) + B_3\,D_i^{1/c_n}\Bigr\}\,. \tag{5.9}$$

Let $D_{(j)}$ denote the jth smallest value of ϵ_i, among indices $i \in \mathcal{I}(x,h)$, and let $V_{(j)}$ be the concomitant value of V_i. (Since the D_i's and V_i's are independent then $V_{(1)}, V_{(2)}, \ldots$ are independent and identically distributed as $V_1, V_2, \ldots$.) Let M denote the least value of m for which there is at least one vector $V_{(i)}$, with $1 \le i \le m$, in each half-sphere. That is, for each unit p-vector v there exist $1 \le i_+(v), i_-(v) \le m$ such that $v^{\mathrm{T}}V_{(i_+(v))} > 0$ and $v^{\mathrm{T}}V_{(i_-(v))} < 0$. Define $M = \infty$ if no such m exists, and let $\mathcal{E}$ denote the event that $M < \infty$. If $\mathcal{E}$ holds then

$$S(x) \equiv \sup_{\gamma_\beta}\,\inf_{1\le i\le M}\,\gamma_\beta^{\mathrm{T}}\,V_{(i)} \le \sup_{\gamma_\beta}\,\min\bigl\{\gamma_\beta^{\mathrm{T}}V_{(i_+(\gamma_\beta))}, \gamma_\beta^{\mathrm{T}}V_{(i_-(\gamma_\beta))}\bigr\} = 0\,.$$

(If $\gamma_\beta \ne 0$ then at least one of the terms within braces is strictly negative, and so the supremum must be attained at $\gamma_\beta = 0$.) More simply, $S(x) \ge 0$, and so $S(x) = 0$. Therefore, by (5.3) and (5.9), if $\mathcal{E}$ holds,

$$\tilde\gamma_2 \le h^2\,S(x) + \tfrac{1}{2}\,h^2\,\|\ddot a\| + \epsilon_{(M)} \le \tfrac{1}{2}\,h^2\,\|\ddot a\| + B_3\,D_{(M)}^{1/c_n}\,,$$

where $\epsilon_{(j)}$ denotes the jth smallest value of ϵ_i. If $\mathcal{E}$ fails then, by the formal definition of $\tilde\gamma_2$, that quantity is infinite, and we have agreed to take $\tilde\gamma_2$ equal to a constant in such cases. Hence, for some $C > 0$,

$$\tilde\gamma_2 \le C\,I(M = \infty) + \tfrac{1}{2}\,h^2\,\|\ddot a\| + B_3\,D_{(M)}^{1/c_n}\,I(M < \infty)\,. \tag{5.10}$$

We may divide the p-variate unit sphere into 2^p regions of equal content, such that, if M_1 denotes the smallest value of m for which each region contains at least one point among $V_{(1)}, \ldots, V_{(m)}$, then $M \le M_1$. Hence, by (5.10),

$$\tilde\gamma_2 \le C\, I(M_1 > N) + \tfrac{1}{2}\, h^2\, \|\ddot a\| + B_3\, D_{(M_1 \wedge N)}^{1/c_n}\,. \tag{5.11}$$

Taking the sequence $V_{(1)}, V_{(2)}, \ldots$ to be unboundedly large rather than stopping at $V_{(N)}$, we may replace $D_{(M_1 \wedge N)}$ on the right-hand side of (5.11) by simply $D_{(M_1)}$, although we keep $I(M_1 > N)$ as it is. It may be proved that for constants $B_5 > 0$ and $B_4 \in (0,1)$, $P(M_1 \ge m) \le B_4\, B_5^m$ whenever $m \ge 1$. Therefore, obtaining the last line using an argument similar to that leading to (5.8),

$$\begin{aligned}
&E\Big\{\big(nh^p\big)^{1/c(x)}\, |\tilde\gamma_2|\Big\}^{2+\eta_1} \\
&\quad \le O(1) + B_3\,\big(nh^p\big)^{(2+\eta_1)/c(x)} \sum_{m=2^p}^{n} E\Big\{I(M_1 = m)\, D_{(m)}^{(2+\eta_1)/c_n}\Big\} \\
&\quad \le O(1) + B_3\,\big(nh^p\big)^{(2+\eta_1)/c(x)} \sum_{m=2^p}^{n} P(M_1 = m)^{1/2}\, \Big\{E D_{(m)}^{2(2+\eta_1)/c_n}\Big\}^{1/2} \\
&\quad \le O(1) + B_6\,\big(nh^p\big)^{(2+\eta_1)/c(x)} \sum_{m=2^p}^{n} B_5^{m/2}\, (m/EN)^{(2+\eta_1)/c_n} = O(1)\,,
\end{aligned}$$

where $0 < B_6 < \infty$. This gives the part of (5.6) pertaining to A_2.

5.2. Proof of Theorem 2. (Recall that we assume that $w_x = 1$.) We shall work with the definition (3.7) of $\check a(x)$. Defining $\tilde\gamma_2(x)$ as at (5.3), and noting (5.4), we have:

$$\check a(x) = \int a(x + h_1 u)\, K(u)\, du + \int \tilde\gamma_2(x + h_1 u)\, K(u)\, du + o_p\big(h^2\big)\,. \tag{5.12}$$

The first integral on the right-hand side, $I_1(x)$ say, equals $a(x) + h^2\, g(x) + o(h^2)$, where $g(x) = \frac{1}{2}\, t^2\, \kappa\, (\nabla^2 a)(x)$, whence it follows that $(nh^p)^{1/c(x)}\, \{I_1(x) - a(x)\} \to \rho\, g(x)$. The stochastic process $S(u) = (nh^p)^{1/c(x)}\, \tilde\gamma_2(x + h_1 u)$ converges weakly to $Q_2(u) \equiv Q_2(u \mid x)$ (see below), whence it follows that the second integral, $I_2(x)$ say, on the right-hand side of (5.12), satisfies $(nh^p)^{1/c(x)}\, I_2(x) \to \int Q_2(u)\, K(u)\, du$.

To appreciate why the finite-dimensional distributions of S converge to those of Q_2, consider the marked point process in $\mathbb{R}^d$, where the ith point is $V_i = h^{-p}(X_i - x)$ and the associated mark is $\zeta_i = \{nh^p\, (t+1)^p\}^{1/c(x)}\, \epsilon_i$. Only the marked points which lie in the disc of radius $t+1$, centred at the origin, contribute to $\check a(x)$, and so we confine attention to those. Define $\zeta_{(1)} < \zeta_{(2)} < \ldots$ to be the ordered values of $\zeta_1 < \zeta_2 < \ldots$, and let $V_{(1)}, V_{(2)}, \ldots$ be the concomitant values of $V_1, V_2, \ldots$. In this new notation, (5.5) continues to hold. From that result it follows, using the argument in the paragraph containing (5.5), that for each finite set $u_1, \ldots, u_k$ in the

sphere of radius $t+1$, centred at the origin, the joint distribution of $S(u_1), \ldots, S(u_k)$ converges to that of $Q_2(u_1), \ldots, Q_2(u_k)$. Tightness of the stochastic process S can be proved using the fact that, defining

$$D(u, j_0) = \sup_{\gamma_\beta} \inf_{1 \le j \le j_0} \left\{ \rho_1 \left(\gamma_\beta^{\mathrm{T}} V_{(i_j(u))} + \tfrac{1}{2} V_{(i_j(u))}^{\mathrm{T}} \ddot{a} V_{(i_j(u))} \right) + (1+p)^{-p/c(x)} \zeta_{(i_j(u))} \right\},$$

where the ordering $j_1(u), j_2(u), \ldots$ is such that $V_{(i_1(u))} < V_{(i_2(u))} < \ldots$ among all indices $i(u)$ such that $\|V_{(i(u))} - u\| \le 1$, the process $D(\,\cdot\,, j_0)$ decreases with increasing j_0.

5.3. Proof of Theorem 3. Derivation of (3.12) is similar to that of the last part of Theorem 1, and so will not be given here. We shall outline proofs of (3.13) and (3.14).

In the case of (3.13), take $a(x) = \delta^2 \psi(x/\delta)$ where $\delta = n^{-1/(p+2c)}$ and ψ is a spherically symmetric function, supported on $\mathcal{V}(0, \frac{1}{2})$, with bounded derivatives of first and second orders, all of them dominated by $\frac{1}{2} C$. Then, $a \in \mathcal{A}$. Consider the problem of discriminating between the models (a) $Y_i = \epsilon_i$ and (b) $Y_i = a(X_i) + \epsilon_i$, using only the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The likelihood-ratio approach, which in view of the Neyman-Pearson lemma is optimal, is to decide in favour of model (b) if and only if the ratio

$$L = \prod_{i=1}^{n} \left[f\{Y_i - a(X_i)\} / f(Y_i) \right]$$

exceeds an appropriate critical point. Here, f is the density at (3.10). If $Y_i = I\, a(X_i) + \epsilon_i$, where $I = 0$ or 1 in cases (a) or (b), respectively, then

$$\log L = (c-1) \sum_{i=1}^{n} \log \left\{ 1 - Y_i^{-1}\, a(X_i) \right\} + \sum_{i=1}^{n} a(X_i)\,.$$

Hence, the likelihood-ratio rule involves deciding in favour of (b) if and only if the sum $\ell = \sum_i \log\{1 - Y_i^{-1}\, a(X_i)\}$ exceeds a critical point.

Asymptotically correct discrimination is readily seen to be impossible if $\nu \equiv n\delta^p$ is bounded; this quantity is of the same order as the number of pairs (X_i, Y_i), for $1 \le i \le n$, that contain information about a. The theorem will follow if we show that, when $\nu \equiv n\delta^p \to \infty$ but $\delta = o(n^{-2/(p+2c)})$ along a subsequence, the probability of correct discrimination using the likelihood-ratio rule, when cases (a) and (b) above both have prior probability $\frac{1}{2}$, converges to $\frac{1}{2}$; it is assumed that all calculations are done for the subsequence.

We may Taylor-expand ℓ, showing that $\ell/\delta^2 = \ell_1 + (I - \frac{1}{2})\, \ell_2 + \ell_3$, where $\ell_1 = -\sum_i \epsilon_i^{-1} \psi_i$, $\ell_2 = \delta^2 \sum_i \epsilon_i^{-2} \psi_i^2$, $\psi_i = \psi(X_i/\delta)$ and, when $\nu \to \infty$ and

$\delta = o(n^{-1/(p+2c)})$, the remainder, ℓ_3, equals $o_p(|\ell_1| + |\ell_2|)$. Using the fact that $0 < c < 2$ it may be proved that $\nu^{-1/c}\,\ell_1$ has a limiting, symmetric, nondegenerate stable distribution with exponent c, and $\delta^{-2}\,\nu^{-2/c}\,\ell_2$ has a limiting, positive, nondegenerate stable law with exponent $c/2$. Therefore, if $\delta = o(n^{-1/(p+2c)})$ then $\ell_2 = o_p(\ell_1)$, from which it follows that the probability of correct classification using the likelihood-ratio rule converges to $\frac{1}{2}$.

To obtain (3.14), let $\mathcal{W}$ denote the cube of diameter 2 inscribed within $\mathcal{V}(0,1)$, with its sides parallel to the coordinate axes. Place into $\mathcal{W}$ a rectangular grid of points, $x_1, \ldots, x_N$ say, with nearest neighbours exactly δ apart and no point distant less than $\frac{1}{2}\delta$ from the boundary of $\mathcal{V}(0,1)$. We may take $N \sim \text{const.}\,\delta^{-p}$ as $\delta \to 0$. Define $a_I(x) = \delta^2 \sum_i I_i\,\psi\{(x - x_i)/\delta\}$, where $I = (I_1, \ldots, I_N)$ is a vector of 0's and 1's. Then $a_I \in \mathcal{A}$ for each choice of I. Since ψ vanishes outside radius $\frac{1}{2}$ from the origin then, for each x, no more than one term in this series is nonzero. Treating the problem of estimating a_I on $\mathcal{R}_h$ as one of discriminating between $I_i = 0$ and $I_i = 1$, for each i such that the sphere of radius $\frac{1}{2}\delta$ centred at x_i intersects $\mathcal{R}_h$, and arguing as in the proof of (3.13), we may derive (3.14).

REFERENCES

AIGNER, D., LOVELL, C.A., KNOX,K. AND SCHMIDT, P. (1977). Formulation and estimation of stochastic frontier production function models. *J. Econometrics* **6**, 21–37.

BREEN, R. (1996). *Regression Models: Censored, Sample Selected, or Truncated Data*. Thousand Oaks, CA: Sage Publications.

CHERNOZHUKOV, V. AND HONG, A. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica* **72**, 1445–1480.

CHRISTENSEN, L. AND GREENE, R. (1976). Economies of scale in U.S. electric power generation. *J. Polit. Econom.* **84**, 653–667.

DONALD, S.G. AND PAARSCH, H.J. (1993). Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *Internat. Econom. Rev.* **34**, 121–148.

DONALD, S.G. AND PAARSCH, H.J. (2002). Superconsistent estimation and inference in structural econometric models using extreme order statistics. *J. Econometrics* **109**, 305–340.

GREENE, W.H. (1990). A Gamma-distributed stochastic frontier model. *J. Econometrics* **46**, 141–163.

HALL, P. (1978). Representations and limit theorems for extreme value distributions. *J. Appl. Probab.* **15**, 639–644.

HALL, P. AND SIMAR. L. (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *J. Amer. Statist. Assoc.* **97**, 523–534.

HARTER, H.L. AND MOORE, A.H. (1965). Maximum-likelihood estimation of the

parameters of gamma and Weilbull populations from complete and from censored samples. *Technometrics* **7**, 639–643.

HILL, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163–1174.

HIRANO, K. AND PORTER, J.R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* **71**, 1307–1338.

JOFRE-BONET, M. AND PESENDORFER, M. (2003). Estimation of a dynamic auction game. *Econometrica* **71**, 1443–1489.

JURECKOVA, J. (2000). Test of tails based on extreme regression quantiles. *Statist. Probab. Lett.* **49**, 53–61.

KNIGHT, K. (2001). Epi-convergence in distribution and stochastic equi-semicontinuity. Manuscript.

KOENKER, R., NG, P. AND PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673–680.

LONG, S.J. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

PAARSCH, H.J. (1992). Deciding between the common and private value paradigms in empirical models of auctions. *J. Econometrics* **51**, 191–215.

PARK, B.U. AND SIMAR, L. (1994). Efficient semiparametric estimation in a stochastic frontier model. *J. Amer. Statist. Assoc.* **89**, 929–936.

PORTNOY, S. AND JUREČKOVÁ, J. (2000). On extreme regression quantiles. *Extremes* **2**, 227–243.

ROBINSON, W.T. AND CHIANG, J.W. (1996). Are Sutton's predictions robust? Empirical insights into advertising, R&D, and concentration. *J. Indust. Economics* **44**, 389–408.

SMITH, R.L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90.

SMITH, R.L. (1994). Nonregular regression. *Biometrika* **81**, 173–183.

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

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

ZIPF, G.K. (1941). *National Unity and Disunity: the Nation as a Bio-Social Organism*. Principia Press, Bloomington, Indiana.

ZIPF, G.K. (1949). *Human Behavior and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley, Cambridge, Mass.

Table 1: *Monte Carlo simulations for* $n = 200$.

Model	a_0	c	Mean(h)	10 Bias	100 Var	100 MSE
1	0.25	0.5	0.048	0.140	0.066	0.086
		1	0.081	1.301	1.014	2.712
		1.5	0.103	3.012	2.091	11.16
	0.50	0.5	0.074	−0.240	0.054	0.112
		1	0.096	0.807	0.667	1.317
		1.5	0.111	2.496	1.987	8.218
2	1	0.5	0.072	−0.052	0.018	0.021
		1	0.091	0.876	0.774	1.540
		1.5	0.108	2.540	2.068	8.528
	2	0.5	0.059	0.047	0.034	0.036
		1	0.086	0.974	1.019	1.968
		1.5	0.107	2.584	2.016	8.694
3	0.25	0.5	0.071	−0.033	0.015	0.016
		1	0.091	0.896	0.866	1.669
		1.5	0.107	2.577	2.102	8.743
	0.50	0.5	0.069	−0.112	0.035	0.047
		1	0.089	0.931	1.044	1.910
		1.5	0.107	2.568	2.030	8.625

Table 2: *Monte Carlo simulations for* $n = 400$.

Model	a_0	c	Mean(h)	10 Bias	100 Var	100 MSE
1	0.25	0.5	0.036	0.009	0.027	0.027
		1	0.067	0.763	0.376	0.958
		1.5	0.081	2.272	0.949	6.112
	0.50	0.5	0.058	−0.273	0.037	0.112
		1	0.091	0.305	0.416	0.509
		1.5	0.101	1.733	0.986	3.990
2	1	0.5	0.056	−0.073	0.008	0.013
		1	0.085	0.478	0.358	0.587
		1.5	0.097	1.820	0.981	4.293
	2	0.5	0.045	−0.013	0.015	0.015
		1	0.079	0.534	0.381	0.666
		1.5	0.095	1.868	0.948	4.436
3	0.25	0.5	0.055	−0.063	0.008	0.012
		1	0.085	0.491	0.362	0.603
		1.5	0.098	1.836	0.998	4.368
	0.50	0.5	0.053	−0.139	0.022	0.041
		1	0.083	0.440	0.403	0.597
		1.5	0.097	1.835	0.989	4.357

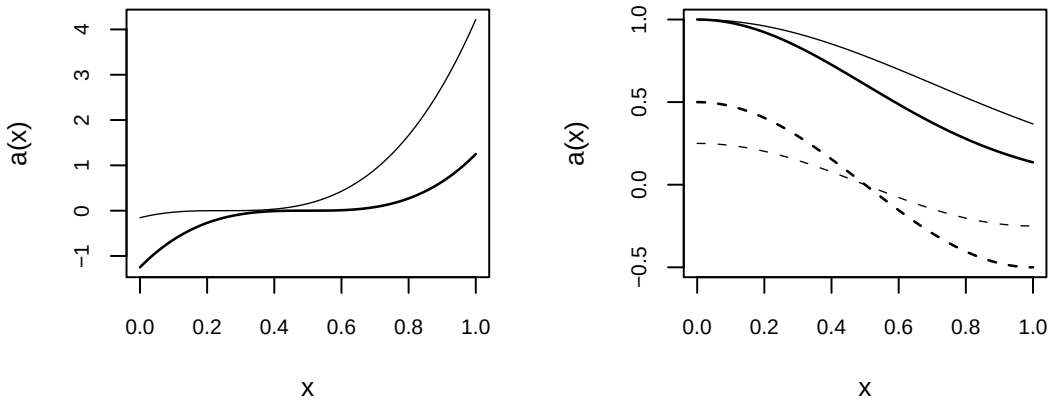

Figure 1: *Graphs of the functions $a(x)$ given in (4.1): the left figure shows $a(x)$ for Model 1 ($a_0 = 0.25$ (thin curve) and $a_0 = 0.50$ (thick curve)), the right figure shows $a(x)$ for Model 2 ($a_0 = 1$ (thin solid curve) and $a_0 = 2$ (thick solid curve)) and Model 3 ($a_0 = 0.25$ (thin dashed curve) and $a_0 = 0.50$ (thick dashed curve)).*

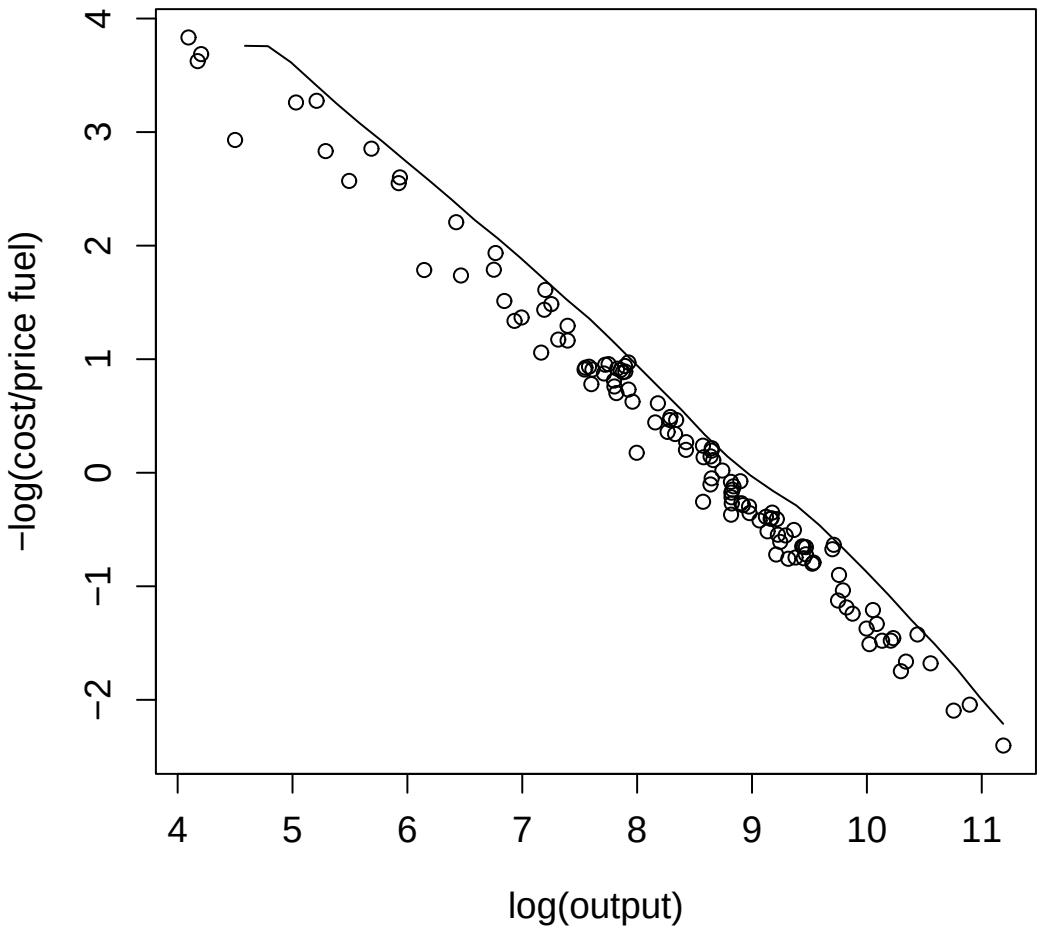

Figure 2: *Scatterplot of the American Electric Utility Data. The observations are represented by circles, the solid curve is the estimated 'regression' (frontier) curve.*