Conditional Ordering Using Nonparametric Expectiles

Yves Aragon, Sandrine Casanova, Ray Chambers and Eve Leconte

Abstract

Expectile regression, and more generally M-quantile regression, can be used to characterise the relationship between a response variable and explanatory variables when the behaviour of "non-average" individuals is of interest. The aim of this paper is to demonstrate how an individual's expectile-order, based on nonparametric estimation of the expectile regression function, can also be used to define a conditional ordering of the individual's value relative to the values of other members of a data set. The relationship between contextual, or "grouping", variables and this ordering can then be investigated. In particular, we propose five estimators of expectile-order, which we compare via simulation. The use of estimated expectile-order to investigate grouping effects is then illustrated using data on physician prescribing behaviour in the Midi-Pyrenees region of France during 1999.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/01

CONDITIONAL ORDERING USING NONPARAMETRIC EXPECTILES

Yves Aragon*, Sandrine Casanova*, Ray Chambers[†] & Eve Leconte*

* *G.R.E.M.A.Q., Université des Sciences Sociales,*
21, allée de Brienne, 31000 TOULOUSE, France
and
L.S.P., Université Paul Sabatier,
118, route de Narbonne, 31062 TOULOUSE Cedex 4, France

[†] *Southampton Statistical Sciences Research Institute*
University of Southampton, Highfield, Southampton SO17 1BJ,
United Kingdom

E-mail : aragon@cict.fr, rc6@soton.ac.uk, leconte@cict.fr,
sandrine.casanova@univ-tlse1.fr

Abstract. Expectile regression, and more generally M-quantile regression, can be used to characterise the relationship between a response variable and explanatory variables when the behaviour of ”non-average” individuals is of interest. The aim of this paper is to demonstrate how an individual’s expectile-order, based on nonparametric estimation of the expectile regression function, can also be used to define a conditional ordering of the individual’s value relative to the values of other members of a data set. The relationship between contextual, or ”grouping”, variables and this ordering can then be investigated. In particular, we propose five estimators of expectile-order, which we compare via simulation. The use of estimated expectile-order to investigate grouping effects is then illustrated using data on physician prescribing behaviour in the Midi-Pyrénées region of France during 1999.

Keywords: conditional expectile, expectile regression, asymmetric regression, local regression, monotonization techniques, order estimation, ordering index

1 Introduction

Regression analysis is a standard tool for modelling the average relationship between a response variable and a set of explanatory variables. Generally this type of analysis models the conditional mean of a response given a set of explanators. However, in some circumstances our interest is not so much in this average relationship, but in an ordering of all individuals based on their ”distance” to the conditional mean. In the following, we investigate an ordering of physicians in the Midi-Pyrénées region of France in 1999. This ordering is with respect to their drug prescribing behaviour and conditions on the characteristics of their practice and other relevant variables. A major problem with constructing such an ordering is that of heteroskedasticity in the regression relationship. In particular, the values associated with individuals whose behaviour deviates from the mean may just reflect heteroskedasticity induced by explanatory variables rather than any intrinsic characteristics of these individuals. Such heteroskedasticity is usually accounted for by a weighted regression fit. However, such an approach typically requires some form of parametric specification for both the regression function and the associated heteroskedasticity, and often assumes errors are symmetrically distributed. There are nonparametric approaches to fitting heteroscedastic models (see Welsh, 1996), but these can be complex. In contrast, we tackle the problem directly by modelling the conditional quantiles of the response given the explanators. Quantiles are part of a general class of distributional location functionals that Breckling and Chambers (1988) refer to as M-quantiles. Beside quantiles, this class contains the expectiles, which generalize the expectation in the same way as quantiles generalize the median (Newey and Powell, 1987), and we base our ordering method on application of this method.

In order to motivate our approach, we consider the problem of monitoring drug prescribing behaviour mentioned above. In particular, let Y be the average value of prescriptions issued by a physician over some fixed period, and assume that a regulatory body (e.g. the Social Security Administration or SSA) has an interest in ranking all physicians in a certain region according to their values of Y. This may be because the SSA wants to identify individual physicians whose prescribing behaviour is substantially different from average prescribing behaviour, or it may be because the SSA is interested in identifying whether there are ”groupings” in these ranks associated with particular sub-regions, indicating inequalities in sub-regional prescription expenditure. In either case, suppose that one assumes that a physician with average prescription value above some threshold, say y_0, generates a ”loss” for the SSA. Then the average loss per physician is:

$$\mathbb{E}((Y - y_0)I(Y > y_0)) \tag{1}$$

while the probability of a physician exceeding this threshold is

$$\mathbb{E}(I(Y > y_0)). \tag{2}$$

Clearly, from an economic point of view, the SSA is more interested in (1) than in (2). Since the value of prescriptions issued by a physician depends on his or her personal characteristics as well as those of the clients of the practice (e.g. their age distribution), the threshold y_0 must also depend on these characteristics.

In practice y_0 is unknown, but we can use the above framework to motivate an approach to ranking individual physicians on the basis of their potential financial risk to the SSD prescription budget. In particular, consider a physician with $Y = y_i$ and $\mathbf{X} = \mathbf{x}_i$. Here $\mathbf{X}$ denotes the (vector-valued) random variable characterising the distribution of physician characteristics across the region of interest. The expected additional loss to the SSD prescription budget caused by an increase in the average value of prescriptions issued by this physician is then

$$\mathbb{E}((Y - y_i)I(Y > y_i) \mid \mathbf{X} = \mathbf{x}_i). \tag{3}$$

A dimension free version of this expected additional loss is obtained by dividing (3) by the average absolute departure from y_i, i.e. $\mathbb{E}(|Y - y_i| \ \mid \mathbf{X} = \mathbf{x}_i)$, leading to the normalized coefficient

$$\frac{\mathbb{E}(|Y - y_i|I(Y > y_i) \mid \mathbf{X} = \mathbf{x}_i)}{\mathbb{E}(|Y - y_i| \ \mid \mathbf{X} = \mathbf{x}_i)}. \tag{4}$$

In particular, the higher the value of this ratio, the lower the financial risk associated with the physician, since the expected loss due to him or her increasing prescription expenditure relative to its current level is greater. In other words, the physician is relatively cheap (in terms of prescription expenditure and after accounting for personal and practice characteristics) as far as the current SSA prescription budget is concerned (Newey and Powell, 1987). Consequently, in order to associate a high ranking with a high risk, we work with the complementary ratio

$$q_i = \frac{\mathbb{E}(|Y - y_i|I(Y \leq y_i)|\mathbf{X} = \mathbf{x}_i)}{\mathbb{E}(|Y - y_i| \ |\mathbf{X} = \mathbf{x}_i)} \tag{5}$$

which we refer to as the *expectile-order* of the physician's prescribing behaviour. The higher the value q_i, the more risky the physician is for the SSA prescription budget. Notice also that (5) parallels the *quantile-order* of the physician's average prescription expenditure, defined by

$$\frac{\mathbb{E}(I(Y \leq y_i)|\mathbf{X} = \mathbf{x}_i)}{\mathbb{E}(1|\mathbf{X} = \mathbf{x}_i)} \tag{6}$$

which corresponds to the probability that a physician with characteristic $\mathbf{x}_i$ has an average prescription expenditure less than or equal to y_i. Since the level of expenditure is of greater interest here than its associated rank, we argue that ranking based on expectile-order is more suitable than ranking based on quantile-order in this situation.

The identity (5) is specific to the realised prescription value y_i. We therefore now generalise the concept of expectile-order so that it applies to arbitrary values of Y and $\mathbf{X}$. In order to do so, we provide a more rigorous definition of expectile regression. Let $F(.|\mathbf{X} = \mathbf{x})$ denote the cumulative distribution function (c.d.f.) of Y given $\mathbf{X} = \mathbf{x}$. Consider the minimization problem

$$\min_{\theta} \int \rho_q(y - \theta) dF(y|\mathbf{X} = \mathbf{x}) \tag{7}$$

where ρ_q is a loss function and q is fixed, $0 < q < 1$. Differentiating the objective function in (7) with respect to θ leads to the estimating equation

$$\int \psi_q(y - \theta) dF(y|\mathbf{X} = \mathbf{x}) = 0 \tag{8}$$

where $\psi_q(u) = \delta\rho_q(u)/\delta u$ is called the influence function. It is well known that if $\psi_q(.)$ equals q for positive values of its argument and equals $-(1-q)$ for negative values of its argument, then the solution to (7) and (8) is the q-quantile of the conditional distribution $F(.|\mathbf{X} = \mathbf{x})$. In contrast, the q-expectile of this conditional distribution is defined by setting

$$\psi_q(u) = \begin{cases} qu & \text{if} \quad u \geq 0 \\ (1-q)u & \text{if} \quad u < 0 \end{cases} \tag{9}$$

in (8). Note that this corresponds to the asymmetric least squares loss function

$$\rho_q(u) = \begin{cases} qu^2 & \text{if} \quad u \geq 0 \\ (1-q)u^2 & \text{if} \quad u < 0. \end{cases} \tag{10}$$

The conditional q-expectile is unique (see Newey and Powell, 1987) and is denoted $m(q, \mathbf{x})$ in what follows. Furthermore, the 0.5-expectile is the expectation of the conditional distribution $F(.|\mathbf{X} = \mathbf{x})$. Substituting the influence function defined by (9) into (8), one obtains a formal definition of $m(q, \mathbf{x})$ as the solution of the equation

$$q = \frac{\mathbb{E}(|Y - m(q, \mathbf{x})| I(Y \leq m(q, \mathbf{x}))|\mathbf{X} = \mathbf{x})}{\mathbb{E}(|Y - m(q, \mathbf{x})| \, |\mathbf{X} = \mathbf{x})}. \tag{11}$$

The general definition of the expectile-order of a sample unit with values $(y_i, \mathbf{x}_i)$ is then the value q_i that satisfies the identity $m(q_i, \mathbf{x}_i) = y_i$.

Newey and Powell (1987) have shown that $m(., \mathbf{x})$ is strictly monotone increasing in q, which guarantees that q can be used to order observations (see e.g. Kokic *et al.*, 1997). Theoretical properties of parametric expectiles are set out in Newey and Powell (1987) and Efron (1991). Breckling and Chambers (1988) extend the concepts of quantile and expectile regression to M-quantile regression and also define a multivariate M-quantile. Yao and Tong (1996) propose a nonparametric estimator of conditional expectiles based on local linear polynomials with a one-dimensional covariate, and establish the asymptotic normality and the uniform consistency of their estimator.

In this paper we focus on the application of expectile-order to the problem of ordering economic performance data, as in Kokic *et al.* (1997). As noted earlier, standard residuals are inadequate in this case because they are sensitive to conditional heteroscedasticity in the data. Instead, we use a nonparametric expectile regression model to estimate the expectile-order. In section 2, we propose five estimators of the expectile-order. The first four require nonparametric estimation of conditional expectiles as a first step, whereas the last one is obtained directly. We compare these estimators using simulated data in section 3. Finally, in section 4, we apply our methods to defining an ordering of a data set containing information about the characteristics and average prescription values of physicians in the Midi-Pyrénées region of France in 1999.

2 Estimation of expectile-orders

In this section we propose five estimators of the expectile-order for the case where the response variable Y is univariate, and the covariate $\mathbf{X}$ is a vector in $\mathbb{R}^p$. Four of the procedures estimate the expectiles $m(q, \mathbf{x})$ on a grid of q values and then, for any given observation, use linear interpolation or logistic smoothing to obtain the corresponding q. The methods are distinguished by the fact that they estimate $m(q, \mathbf{x})$ by a locally constant Nadaraya-Watson kernel estimator, a locally linear kernel estimator, a locally linear mean preserving monotone kernel estimator and a locally linear isotonic regression kernel estimator. The fifth estimates the expectile-order directly. The observed sample values are denoted $(Y_i, \mathbf{X}_i)_{i=1}^n$ in what follows.

2.1 Expectile-order based on locally constant expectile regression

A kernel-based estimator $m_{LC}(q, \mathbf{x})$ of $m(q, \mathbf{x})$ that is equivalent to fitting a local constant to this function is the solution to the minimization problem

$$\min_{\theta \in R} \int \rho_q(y - \theta) d\hat{F}_n(y|\mathbf{X} = \mathbf{x}) \tag{12}$$

where

$$\hat{F}_n(y|\mathbf{X}=\mathbf{x}) = \frac{\sum_{i=1}^{n} K(\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}) I(Y_i \leq y)}{\sum_{i=1}^{n} K(\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}})}$$

is the Nadaraya-Watson kernel estimator of the conditional c.d.f. $F(.|\mathbf{X}=\mathbf{x})$, K is a multivariate kernel function, **h** is a vector of suitable bandwidths and the loss function ρ_q is defined by (10). Differentiating (12) with respect to θ leads to the estimating equation

$$\sum_{i=1}^{n} \psi_q(Y_i-\theta) K(\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}) = 0 \tag{13}$$

with ψ_q as in (9). Defining $V_{q,i}(\mathbf{x}) = (I_i - 2qI_i + q)K(\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}})$ where $I_i = I(Y_i \leq \theta)$ and solving (13) leads to the estimator $m_{LC}(q,\mathbf{x})$, which can be written as a weighted average of the sample values of Y,

$$m_{LC}(q,\mathbf{x}) = \frac{\sum_{i=1}^{n} V_{q,i}(\mathbf{x}) Y_i}{\sum_{i=1}^{n} V_{q,i}(\mathbf{x})}. \tag{14}$$

For general q the estimator $m_{LC}(q,\mathbf{x})$ can only to be computed iteratively since $V_{q,i}(\mathbf{x})$ depends on θ. This estimator is strictly monotone increasing in q, so that an estimator $q_{LC}(y,\mathbf{x})$ of the expectile-order of an observation with $Y=y$ can be directly computed by linear interpolation over a grid of values of q defined for each value of **x**. That is, if q_L and q_U are the two adjacent values on this grid such that $m_{LC}(q_L,\mathbf{x}) < y < m_{LC}(q_U,\mathbf{x})$ then the estimated expectile-order of a sample unit with values y and **x** is $q(y,\mathbf{x}) = \alpha(y,\mathbf{x})q_L + (1-\alpha(y,\mathbf{x}))q_U$ where

$$\alpha(y,\mathbf{x}) = \frac{m_{LC}(q_U,\mathbf{x}) - y}{m_{LC}(q_U,\mathbf{x}) - m_{LC}(q_L,\mathbf{x})}. \tag{15}$$

2.2 Expectile-orders based on local linear estimators

Alternatively we consider nonparametric estimation of the expectile regression function based on a kernel weighted local linear fit (Yao and Tong, 1996). Given a $p \times 1$ vector **u** we define $\mathbf{u}^{*'} = [1\ \mathbf{u}']$. A locally linear nonparametric estimator of $m(q,\mathbf{x})$ is then

$$m_{LL}(q,\mathbf{x}) = \mathbf{x}^{*'} \hat{\beta}_q(\mathbf{x}) \tag{16}$$

where $\hat{\beta}_q(\mathbf{x})$ is the solution to the minimization problem

$$\min_{\mathbf{b} \in R^{p+1}} \int \rho_q(y - \mathbf{x}^{*'}\mathbf{b}) d\hat{F}_n(y|\mathbf{x}). \tag{17}$$

Differentiating (17) with respect to **b** leads to the estimating equation

$$\sum_{i=1}^{n} \psi_q(Y_i - \mathbf{X}_i^{*'}\mathbf{b}) K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}}) \mathbf{X}_i^{*'} = 0. \tag{18}$$

Let $\mathbf{Y}$ be the $n \times 1$ vector of sample data for the response variable, $X^{*'} = [\mathbf{X}_1^{*'} \cdots \mathbf{X}_n^{*'}]$ with $\mathbf{X}_i^{*'}$ defined similarly as $\mathbf{u}^{*'}$ and let $V_q(\mathbf{x})$ be the $n \times n$ diagonal matrix of weights $\{V_{q,i}(\mathbf{x})\}$, where the $V_{q,i}(\mathbf{x})$ were defined in the previous section. The solution to (18) is then

$$\hat{\beta}_q(\mathbf{x}) = (X^{*'} V_q(\mathbf{x}) X^*)^{-1} X^{*'} V_q(\mathbf{x}) \mathbf{Y}.$$

Note that $\hat{\beta}_q(\mathbf{x})$ must also be computed iteratively since the matrix $V_q(\mathbf{x})$ depends on **b**.

The estimator $m_{LL}(q, \mathbf{x})$ is not necessarily a nondecreasing function of q at every value of **x**. That is, the fitted expectile surfaces obtained by solving (18) can cross in the sample **x**-data range. This problem is also discussed in Kokic *et al.* (1997). He (1997) describes a restricted version of quantile regression that avoids such crossing. Craig and Ng (2001) encounter the same problem when using smoothing splines to estimate conditional quantiles in an analysis aimed at identifying employment subcenters in a multicentric urban area. Here we tackle this problem by constraining the estimator $m_{LL}(q, \mathbf{x})$ so that it is monotone with respect to the values q on a grid Q defined at every sample value of **x**. In particular we adapt the technique of Mukarjee and Stern (1994) so that, for q in Q, the estimator $m_{LL}(q, \mathbf{x})$ is replaced by the mean preserving monotone estimator $m_{MPM}(q, \mathbf{x})$

$$m_{MPM}(q, \mathbf{x}) = \begin{cases} \min_{q' \in Q,\ q \leq q' \leq 0.5} m_{LL}(q', \mathbf{x}) & \text{if} \quad q \in]0, 0.5] \\ \max_{q' \in Q,\ 0.5 \leq q' \leq q} m_{LL}(q', \mathbf{x}) & \text{if} \quad q \in]0.5, 1[\end{cases}$$

An alternative approach is to use isotonic regression (Robertson *et al.*, 1998) to construct a monotone estimator of $m(q, x)$. This leads to the estimator $m_{IRM}(q, \mathbf{x})$, which is the nearest monotone estimator of $m(q, x)$ according to the L_2 norm. Let $Q = \{q_1, \cdots, q_s\}$ be the grid of values of q with $q_1 \leq \cdots \leq q_s$. Then for q_i in Q, $m_{IRM}(q_i, x)$ is defined by

$$m_{IRM}(q_i, \mathbf{x}) = \min_{\{i \leq t\}} \max_{\{r \leq i\}} Av\{m_{LL}(q_k, \mathbf{x}), r \leq k \leq t\},$$

where $Av(X_1, \cdots, X_m)$ is the empirical mean of the sequence $X_1, \cdots, X_m$. For both methods of monotonization, the estimated expectile-order of each observation $(y, \mathbf{x})$ is then calculated by linear interpolation as in (15), leading to two estimators of q that we denote by $q_{MPM}(y, \mathbf{x})$ and $q_{IRM}(y, \mathbf{x})$ respectively.

Finally, as an alternative to direct monotonization of the conditional expectiles, one can fit a linear model to the logits of the values in the grid Q using the estimated conditional expectile values $m_{LL}(q, \mathbf{x})$ calculated on this grid at a fixed value $\mathbf{x}$ as explanators. The estimated expectile-order $q_{LR}(y, \mathbf{x})$ for a point $(y, \mathbf{x})$ is then obtained as the predicted value generated by this model at the value y.

2.3 A direct estimator of the expectile-order

From (11) we see that the value y of a data point $(y, \mathbf{x})$ is the expectile $m(q, \mathbf{x})$ where

$$q = \frac{\mathbb{E}(|Y - y| I(Y \leq y) | \mathbf{X} = \mathbf{x})}{\mathbb{E}(|Y - y| | \mathbf{X} = \mathbf{x})}. \tag{19}$$

For each $(y, \mathbf{x})$, we can estimate the numerator and the denominator of (19) using weighted Nadaraya-Watson type kernel estimators (Hall *et al.*, 1999). The resulting estimator of the expectile-order is then

$$q_{ALNW}(y, \mathbf{x}) = \frac{\sum_{i=1}^{n} |Y_i - y| I(Y_i \leq y) K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}}) w_i(\mathbf{x})}{\sum_{i=1}^{n} |Y_i - y| K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}}) w_i(\mathbf{x})}, \tag{20}$$

where the $w_i(\mathbf{x})$'s define a set of calibrating weights, i.e. they satisfy $w_i \geq 0$, $\sum_i w_i = 1$ and

$$\sum_i (\mathbf{X}_i - \mathbf{x}) K(\frac{\mathbf{X}_i - \mathbf{x}}{\mathbf{h}}) w_i(\mathbf{x}) = 0.$$

The above constraints do not uniquely define the $w_i(\mathbf{x})$'s, and so we calculate these weights by minimising $\sum_i w_i^2$ subject to these constraints. This ensures that w_i stays close to $\frac{1}{n}$. Put $\mathbf{u}_i(\mathbf{x}) = (\mathbf{X}_i - \mathbf{x}) K(\frac{\mathbf{X}_i - \mathbf{x}}{\mathbf{h}})$. The $p \times n$ matrix $U(\mathbf{x})$ is then defined by $U(\mathbf{x}) = (\mathbf{u}_1(\mathbf{x})\ \mathbf{u}_2(\mathbf{x})\ \cdots \mathbf{u}_n(\mathbf{x}))$ with $\bar{\mathbf{U}}(\mathbf{x}) \in \mathbb{R}^p$ the mean vector of the rows of U. Straightforward calculation yields

$$(w_1(\mathbf{x})\ w_2(\mathbf{x}) \cdots w_n(\mathbf{x}))' = \frac{1}{n} \mathbf{1}_n - \frac{|A(\mathbf{x})|}{|B(\mathbf{x})|} (U(\mathbf{x}) - \bar{\mathbf{U}}(\mathbf{x}) \mathbf{1}_n')' A^{-1}(\mathbf{x}) \bar{\mathbf{U}}(\mathbf{x})$$

where $A(\mathbf{x}) = U(\mathbf{x})U'(\mathbf{x})$ and $B(\mathbf{x}) = (U(\mathbf{x}) - \bar{\mathbf{U}}(\mathbf{x})\mathbf{1}'_n)(U(\mathbf{x}) - \bar{\mathbf{U}}(\mathbf{x})\mathbf{1}'_n)'$ are $p \times p$ matrices. Note that Hall *et al.* (1999) define the weights w_i so that they maximize $\prod_i w_i$, i.e. these authors seek to minimise the Kullback distance of $\{w_i\}$ from $\frac{1}{n}$. Unfortunately, we experienced convergence problems when attempting to apply this criterion. Furthermore, $q_{ALNW}(y, \mathbf{x})$ is a nondecreasing function of y because (20) is equal to (19) when the conditional distribution function $F(y|\mathbf{X} = \mathbf{x})$ is

$$\frac{\sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{X}_i)}{\mathbf{h}})w_i(\mathbf{x})I(Y_i \leq y)}{\sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}})w_i(\mathbf{x})}.$$

Using the results in Hall *et al* (1999), it can be shown that the numerator and the denominator of q_{ALNW} (both divided by $\sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}})w_i(\mathbf{x})$) are local linear estimators in which the weights are $K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}})w_i(\mathbf{x})$ instead of $K(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{h}})$. Furthermore, under suitable regularity conditions, these estimators are first order equivalent to classical local linear estimators. Finally, we observe that since the computation of (20) is very fast, an estimator $m_{ALNW}(q, \mathbf{x})$ of $m(q, \mathbf{x})$ can be derived as follows. We first calculate the estimated expectile-orders $q_{ALNW}(y, \mathbf{x})$ on a very fine grid of y values. Then, for a given value q and a fixed value of the covariate $\mathbf{x}$, $m_{ALNW}(q, \mathbf{x})$ is obtained by linear interpolation.

3 A simulation study

3.1 Description

In this section we investigate the finite sample performance of the five estimators of the expectile regression functions that were defined in the previous section, as well as their corresponding estimators of the expectile-orders of the sample values. Data values for $S = 500$ samples, each of size $n = 200$ were simulated, with the covariate X defined as the sum of two independent variables uniformly distributed on $[0, 2.5]$ and the value of Y given $X = x$ drawn from a Gaussian distribution with mean $m(x) = 20 + (0.8x - 2)^3$ and standard deviation 2.5. With this definition, the corresponding conditional q-expectile of Y at $X = x$ is $m(q, x) = m(x) + 2.5\ e_q$, where e_q is the q-expectile of a standard Gaussian distribution. All kernel-based estimators used the Epanechnikov kernel. We chose three bandwidths (one for each of the estimators m_{LC}, m_{LL} and q_{ALNW}) using three separate cross validation exercises. Since the choice of bandwidth precedes monotonization, the estimators m_{MPM} and m_{IRM} used the same bandwidth as m_{LL}. Bandwidth choice for the estimators m_{LC} and m_{LL} was based on extending the classical least squares cross-validation technique to the case of expectiles, with the selected bandwidth minimizing

$$\sum_{i=1}^{n} \rho_q(Y_i - m_{EST,-i}(q, X_i))$$

on a grid of 20 regularly spaced bandwidth values in [0.8,5] (the length of this interval roughly corresponds to the range of the covariate). Here *EST* denotes the type of smoother used (LC or LL) and $m_{EST,-i}$ is calculated using the data set $\{(y_j, x_j),\ j \neq i\}$.

A cross-validation criterion was also used to determine the bandwidth for the direct estimator of expectile-order. For a given observation (y_i, x_i), we define the random variables

$$Y_{1i} = |Y - y_i| I(Y \leq y_i) \text{ and } Y_{2i} = |Y - y_i|.$$

Let $(Y_{1ij}, X_j)_{j=1}^n$ and $(Y_{2ij}, X_j)_{j=1}^n$ be the observed sample data. Let $m_{1i}(x) = \mathbb{E}(Y_{1i}|X = x)$ and $m_{2i}(x) = \mathbb{E}(Y_{2i}|X = x)$. The true expectile-order of the observation (y_i, x_i) is then $q(y_i, x_i) = \dfrac{m_{1i}(x_i)}{m_{2i}(x_i)}$ and the estimator $q_{ANLW}(y_i, x_i)$ is $\dfrac{\hat{m}_{1i}(x_i)}{\hat{m}_{2i}(x_i)}$ where $\hat{m}_{ki}(x_i)$, $k = 1, 2$, is the weighted Nadaraya-Watson estimator of the conditional mean $m_{ki}(x_i)$. Optimal bandwidths for the $\hat{m}_{ki}(x_i)$, $k = 1, 2$, $i = 1, \cdots, n$ are then obtained by minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (Y_{kij} - \hat{m}_{ki,-j}(x_i))^2,\ k = 1, 2 \tag{21}$$

where $\widehat{m}_{ki,-j}$, $k = 1, 2$, is calculated using the data set $\{(y_l, x_l),\ l \neq j\}$. A Taylor expansion of q_{ALNW} in the neighborhood of $(m_{1i}(x_i), m_{2i}(x_i))$ leads to the approximation

$$q_{ALNW}(y_i, x_i) \simeq \frac{m_{1i}(x_i)}{m_{2i}(x_i)} + a_i(\hat{m}_{1i}(x_i) + b_i \hat{m}_{2i}(x_i))$$

with $a_i = \dfrac{1}{m_{2i}(x_i)}$ and $b_i = -\dfrac{m_{1i}(x_i)}{m_{2i}(x_i)}$. The same bandwidth is then used in both numerator and denominator of q_{ALNW}, and is chosen so that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \{(Y_{1ij} - \hat{m}_{1i,-j}(x_i)) + b_i(Y_{2ij} - \hat{m}_{2i,-j}(x_i))\}^2$$

is minimized. The coefficients b_i in this expression are estimated using the component specific optimal bandwidths determined by minimizing (21).

The estimators $m_{LC}(q,x)$, $m_{LL}(q,x)$, $m_{MPM}(q,x)$, $m_{IRM}(q,x)$ and $m_{ALNW}(q,x)$ of the conditional expectile function were then computed for a set of $M = 49$ regularly spaced values $\{x_1, \cdots, x_M\}$ of x in $[0.1, 4.9]$ and for a grid of $L = 9$ values of q, corresponding to $Q = \{.01, .05, .1, .2, .5, .8, .9, .95, .99\}$. Since we know the true conditional expectile function, the mean squared error (MSE) of each estimator $m_{EST}(q,x)$ of $m(q,x)$ can be evaluated as

$$\text{MSE}(m_{EST}, q, x) = \frac{1}{S}\sum_{s=1}^{S}(m_{EST_s}(q,x) - m(q,x))^2,$$

where m_{EST_s} denotes the estimator of m for the sth sample. We also compute the mean averaged squared error (MASE) defined by

$$\text{MASE}(m_{EST}, q) = \frac{1}{SM}\sum_{s=1}^{S}\sum_{m=1}^{M}(m_{EST_s}(q,x_m) - m(q,x_m))^2.$$

For EST in the set $\{LC, MPM, IRM, LR, ALNW\}$, the performance of an estimator $q_{EST}(y,x)$ of the expectile-order of a data value (y,x), based on corresponding estimated conditional expectile functions at each value q in the grid Q, is then evaluated by calculating its mean absolute deviation error (MADE) for each sample s (see Hall *et al.*, 1999),

$$MADE(q_{EST_s}) = \frac{1}{LM}\sum_{l=1}^{L}\sum_{m=1}^{M}|q_{EST_s}(y_{lm}, x_m) - q_l|,\ s = 1, \cdots, S,$$

where y_{lm} satisfies $m(q_l, x_m) = y_{lm}$.

3.2 Results

3.2.1 Estimators of conditional expectile functions

Table 1 shows the values of MASE for q in Q. Notice that the estimator m_{LL} performs better than the estimator m_{LC} and that monotonization leads to an improvement in MASE. The monotonized estimators m_{MPM} and m_{IRM} have similar performances, with m_{MPM} performing better for extreme values of q and m_{IRM} performing better for values of q close to $q = 0.5$. The estimator m_{ALNW} performs best for extreme values of q, but is inefficient for intermediate values.

q	m_{LL}	m_{MPM}	m_{IRM}	m_{LC}	m_{ALNW}
.01	1.8239	1.7425	1.7781	2.6970	0.8712
.05	1.2641	1.1527	1.1787	1.9860	0.9757
.1	.84825	.82594	.82987	1.3145	0.9313
.2	.71979	.71043	.69274	1.1940	0.8718
.5	.61578	.61578	.59281	1.1305	0.8718
.8	.71693	.71042	.69471	1.1924	0.8917
.9	.88439	.85448	.84264	1.3061	0.9562
.95	1.2637	1.1829	1.1906	2.0011	1.0164
.99	1.8681	1.8054	1.8239	2.7129	0.9126

Table 1: Values of MASE for estimators of conditional expectiles at the values of q in Q.

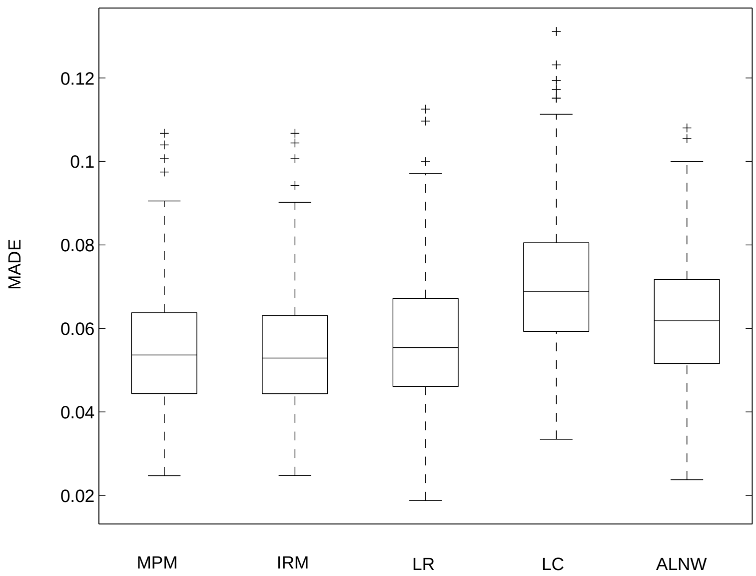

Figure 1: Boxplots of the MADE values of the estimated conditional expectile-orders generated by the five methods for $S = 500$ samples. The corresponding means are 0.0545, 0.0544, 0.0575, 0.0708 and 0.0624.

3.2.2 Estimators of expectile-orders

Boxplots of MADE for the five estimators q_{MPM}, q_{IRM}, q_{LC}, q_{LR} and q_{ALNW} of expectile-orders are shown in Figure 1. As with estimation of conditional expectiles, the estimators q_{MPM} and q_{IRM} based on local linear regression perform better than the estimator q_{LC} based on locally constant regression and the direct estimator q_{ALNW}. The median MADE value for the estimator q_{LR} is only marginally higher than the median MADE values of q_{MPM} and q_{IRM}. However, its variability is greater. On the basis of these rather limited simulation results, it appears that the expectile-order estimators q_{MPM} and q_{IRM} based on monotonized expectile fits may be preferable.

4 An application

4.1 The data set

We focus on a data set that contains measurements on 2801 physicians in the Midi-Pyrénées region of France during 1999, including most of the general practice physicians in this region. The study variable, denoted Y, measures the drug prescribing activity of a physician, and is defined as the logarithm of the ratio of the value of drug prescriptions issued by the physician over the year divided by the number of ”acts” carried out by the physician over the same period. An act may be a house call or a consultation. In addition to this variable, the data set contains a number of indicators of a physician's practice and activity characteristics as well as the physician's age and gender. These variables are denoted $X_1, \cdots, X_{15}$ and are listed in Table 2. Each physician works in a canton (a small county). For each canton we also have demographic statistics and other characteristics, e.g. level of education and level of unemployment. These variables are denoted $Z_1, \cdots, Z_{11}$ and are listed in Table 3. We do not have direct measures of the health status of the patients for whom the prescriptions are issued. Two levels of explanatory variables are thus available - physician level and canton level. We use these data to quantify the prescribing performance of a physician. In particular we calculate each physician's expectile-order based on the the physician's value of drug prescription per act, given his or her characteristics, including practice characteristics. Our aim is to investigate the extent to which variation in these expectile-orders can be ”explained” using the canton-level variables defined in Table 3.

Y	Logarithm of the value of prescriptions per act
X_1	Physician seniority (years)
X_2	Total practice size
X_3	% of practice less than 16
X_4	% of practice from 60 to 69
X_5	% of practice more than 70
X_6	% of practice who don't pay medical fees
X_7	% of practice who are farm employed
X_8	% of practice who are self employed
X_9	Number of consultations and house calls
X_{10}	Proportion of house calls
X_{11}	Number of consultations per patient
X_{12}	Number of house calls per patient
X_{13}	Average fee per patient
X_{14}	Age of physician
X_{15}	Gender of physician

Table 2: Physician and practice variables.

4.2 Dimension reduction

Nonparametric regression can become unstable if there are many covariates. Since the ordering methodology described in this paper depends on a predictive, rather than interpretative, regression model, it is advisable to reduce the dimension of the covariate space by taking into account the dependence between the covariates and the response variable. This can be done through a Sliced Inverse Regression (SIR) analysis (Li, 1991, and Cook, 1994, 1996). This method is a fast exploratory analysis tool producing a small number of synthetic indices (linear combinations of the covariates). Nonparametric regression modelling then proceeds using these indices as covariates. A SIR of the response variable based on the physician and practice variables in table 2 gives 6 major eigenvalues (see Table 4).

These eigenvalues fall sharply after the second eigenvalue. Consequently we use the first two SIR indices as covariates in the nonparametric regression fit to the expectiles of the value of drug prescription per act. These indices are denoted $EDR1$ and $EDR2$ in what follows. Table 5 shows the correlations between these two indices and the variables $Y, X_1, \cdots, X_{15}$ used in the SIR. The dependent variable appears first.

It can be seen that both indices are highly associated with the proportion of house calls and the number of house calls per patient. $EDR1$ is also highly asso-

Z_1	Mean income per capita (1996)
Z_2	Density of population
Z_3	% of population less than 15
Z_4	% of population from 60 to 69
Z_5	% of population more than 70
Z_6	Number of deaths per 1000 inhabitants
Z_7	Number of births per 1000 inhabitants
Z_8	Retirement rate (in %)
Z_9	Unemployment rate (in %)
Z_{10}	Employment rate (in %)
Z_{11}	Number of physicians per 1000 inhabitants

Table 3: Canton variables.

0.3206	0.1841	0.0330	0.0203	0.0134	0.0107

Table 4: Eigenvalues of SIR

ciated with the level of activity of the physician, the percentage of old persons in the practice and the average fee per patient. In contrast $EDR2$ is highly associated with the percentage of young people in the practice and the percentage of the practice aged from 60 to 69.

In an effort to improve the estimation of these expectiles and of the consequent expectile-orders, we also investigated bringing the canton variables in Table 3 into the regression model. Here we performed a SIR of the amount of drug prescription per act on the combined set of variables $X_1, \cdots, X_{15}, Z_1 \cdots, Z_{11}$, with values of the variables $Z_1, \cdots, Z_{11}$ replicated for each physician in a canton. From an inspection of the resulting eigenvalues we again decided to retain two indices. Both were highly correlated with the corresponding indices identified from the first SIR (correlations of .984 and .959 respectively). Consequently, the introduction of canton level effects did not lead to any real change in the SIR indices, and so we proceeded to estimate the expectile-orders of the physicians in our data set conditioning only on the SIR indices $EDR1$ and $EDR2$ based on Y and $X_1, \cdots, X_{15}$.

Variable	$EDR1$	$EDR2$
Y	0.542	-0.015
X_1	0.203	0.000
X_2	0.409	0.015
X_3	0.040	0.676
X_4	0.311	-0.709
X_5	0.567	-0.429
X_6	0.348	-0.140
X_7	0.280	-0.061
X_8	-0.099	-0.118
X_9	0.569	0.271
X_{10}	0.636	0.163
X_{11}	-0.075	0.360
X_{12}	0.582	0.309
X_{13}	-0.066	0.029
X_{14}	0.078	-0.124
X_{15}	-0.251	-0.060

Table 5: Correlations between SIR indices and physician and practice variables.

4.3 Measuring the quality of an expectile fit

In standard linear regression, the adjusted coefficient of determination is used to measure the quality of the regression fit, with a low value of this coefficient indicating low explanatory power or the presence of misspecification. To avoid misspecification issues, we use local regression techniques to estimate conditional expectiles. Replacing the 'square' function by the loss function ρ_q defined by (10), we adapted the adjusted coefficient of determination to the case of local expectile regression, leading to the coefficient

$$R_q^2 = 1 - \frac{\sum_{i=1}^{n} \rho_q(y_i - m_{EST}(q, \mathbf{x}_i))/(n - \nu(q))}{\sum_{i=1}^{n} \rho_q(y_i - \hat{m}(q))/(n-1)}.$$

Here $\hat{m}(q)$ denotes the unconditional empirical q-expectile of Y, that is the value of θ that minimizes $\sum_{i=1}^{n} \rho_q(y_i - \theta)$, and EST belongs to $\{LL, LC\}$. The local regression estimator $m_{EST}(q, \mathbf{x})$ is linear in y, and so for each x can be written

$m_{EST}(q, \mathbf{x}) = \sum_{i=1}^{n} l_i(q, \mathbf{x}) y_i$ (see Loader, 1999). As in linear regression, the constant $\nu(q)$ is therefore defined as the trace of the matrix $L(q) = [l_j(q, \mathbf{x}_i)]_{i=1,\cdots,n}^{j=1,\cdots,n}$.

By definition R_q^2 is a global measure of the quality of the local expectile regression of order q. As with the usual coefficient of determination, a low value of R_q^2 indicates low dependence of Y on **X**, so that the conditional distribution of Y is not well described by **X**. In such a situation the resulting expectile-order estimates will not be reliable. Notice that R_q^2 can also be used as a model-selection tool.

4.4 Expectile modelling of the physician and practice variables

We estimated the expectile-orders of the physicians using the estimator q_{MPM} described in section 2, with an interpolation grid $Q = \{.01, .1, .2, .5, .8, .9, .99\}$. As in the simulations, we used a locally linear smoother with a bivariate Epanechnikov kernel with bandwidths set to 20% of the range of each SIR index. Figure 2 is the histogram of the resulting estimated expectile-orders. Physicians with estimated expectile-orders in the tails of this distribution can be considered to have displayed extreme prescribing behaviour (in both a negative as well as a positive sense) relative to physicians with similar characteristics in the Midi-Pyrénées region in 1999. Note that the ”high cost” physicians are more numerous than the ”low cost” physicians. This can be contrasted with quantile orders, which are necessarily uniformly distributed.

A question of some interest is the extent to which the variation in expectile-orders of individual physicians can be explained by canton-level effects. The presence of such effects in these estimated expectile-orders can be seen in Figure 3. This shows the boxplots of estimated expectile-orders for the 12 larger cantons. Note that the median of these orders varies significantly between cantons. Thus, for the canton of Rodez, a rich rural canton, the median of these orders is close to 0.8 whereas for the canton of Auch, the median is just above 0.4. For the canton of Toulouse, the main city of the Midi-Pyrénées region, the median is near 0.6. An analysis of variance of the logit of the expectile-orders with respect to the canton variable indicates that the average value of the estimated expectile-order varies significantly between cantons (p = 0.030).

Finally, in Table 6 we show the values of the R_q^2 coefficient for different values of q and two sets of explanatory variables: the first where the nonparametric regression fit is carried out using only the first SIR index $EDR1$ and the second one where this fit is based on both $EDR1$ and $EDR2$. Notice that taking $EDR2$ into account improves the fit at each value of q. Notice also that R_q^2 is a decreas-

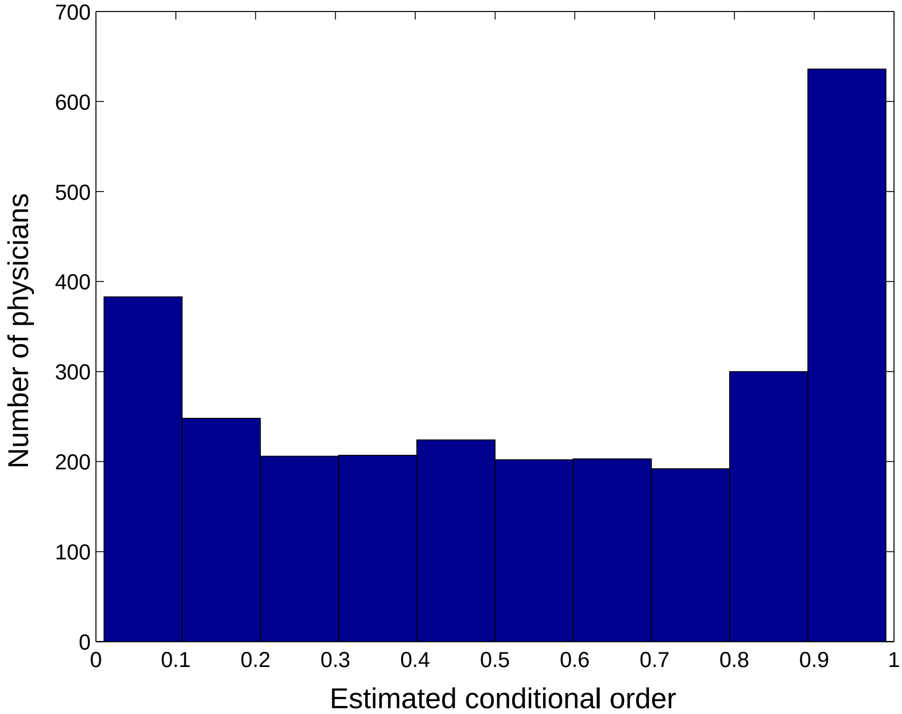

Figure 2: Histogram of the estimated expectile-orders for all 2801physicians.

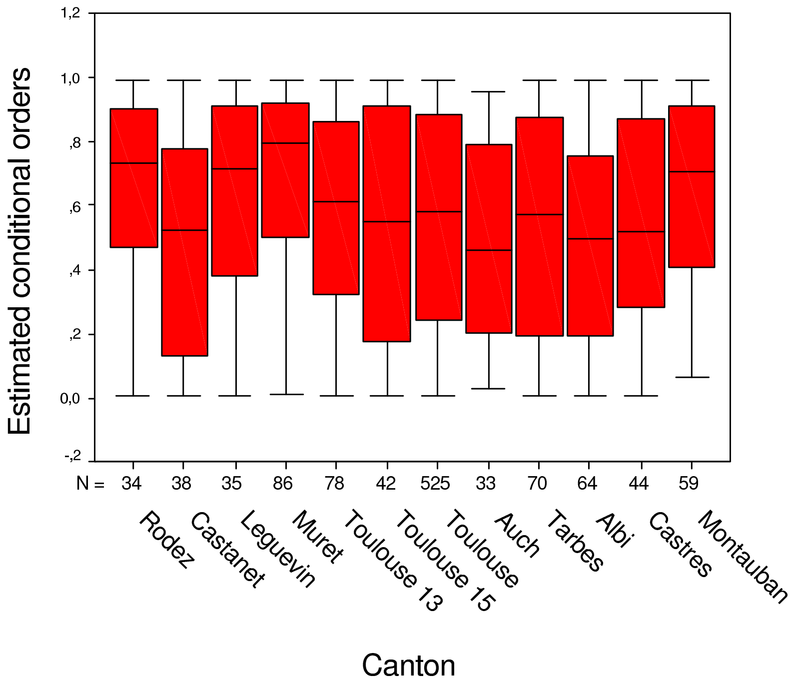

Figure 3: Boxplots of the estimated conditional expectile-orders for the 12 larger cantons.

q	.01	.1	.2	.5	.8	.9	.99
$EDR1$	0.58757	0.48200	0.42672	0.31870	0.21071	0.15333	0.04549
$EDR1$ and $EDR2$	0.68036	0.54896	0.48168	0.35914	0.26604	0.23798	0.22776

Table 6: Values of adjusted R_q^2 coefficient for one and two SIR indices and for different values of q.

ing function of q in Table 6. Justification for this behaviour can be seen in Figure 4. This shows that, conditionally on $EDR1$, small values of Y (corresponding to small expectile-orders) are more sensitive to variation in $EDR1$ than large values of Y.

The conditional expectile-orders of physicians in the Midi-Pyrénées region were also estimated directly using q_{ALNW}. Computation of this estimator is extremely fast (typically 1000 times quicker than for the estimator q_{MPM}). A scatterplot of q_{ALNW} versus q_{MPM} (see Figure 5) shows that these estimators coincide for most physicians in the data set, with a correlation of 0.99. Note that direct estimation of the expectile-order is appropriate when comparison of sample individuals is of primary interest. On the other hand, estimators of the expectiles curves may be useful when a global description of the conditional distribution is required.

5 Discussion

In this paper we introduce the concept of the expectile-order of an observation and show how it can be estimated via nonparametric expectile regression. We also demonstrate its application in the context of an analysis of the prescribing behaviour of a population of physicians. In particular, we show how the relationship between these expectile-orders and contextual variables (e.g. cantonal affiliation) can be easily tested. In this context our approach can be seen as offering a nonparametric alternative to more standard multilevel parametric modelling of data with group structure. Finally, we note that all the ideas presented in this paper can be generalized to standard quantiles, and more generally to M-quantiles (Breckling and Chambers, 1988). Such generalizations offer the promise of orderings that are robust to outlying values in Y (since they are based on bounded influence functions). However they also lack the interpretability of expectile-ordering, in the sense that they do not rank on the basis of expected loss.

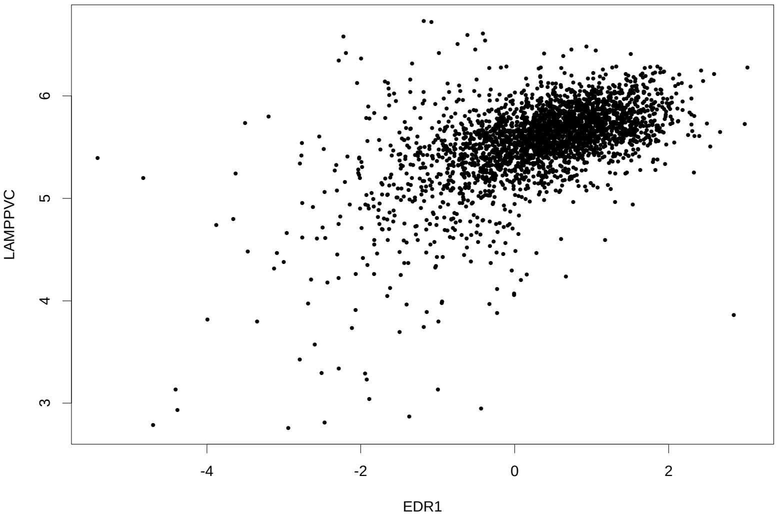

Figure 4: Plot of Y vs. $EDR1$.

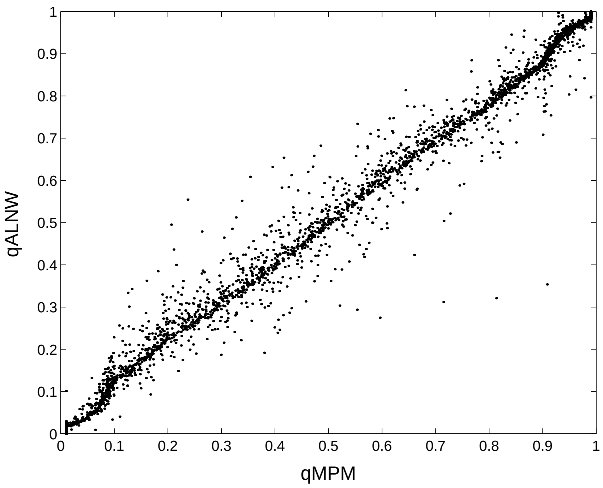

Figure 5: Plot of q_{ALNW} vs. q_{MPM}.

References

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761-771.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-189.

Cook, R. D. (1996). Graphics for regression with a binary response. *Journal of the American Statistical Association*, **91**, 983-992.

Craig, S.G. and Ng, P.T. (2001). Using quantile smoothing splines to Identify employment subcenters in a multicentric urban area. *Journal of Urban Economics*, **49**, 100-120.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, **1**, 93-125.

Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154-163.

He, X. (1997). Quantile curves without crossing. *American Statistician*, **51**, 186-192.

Kokic, P., Chambers, R., Breckling, J. and Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics*, **15**, 445-451.

Li K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-342.

Loader, C. (1999). *Local Likelihood and Regression*, Springer.

Mukarjee, H. and Stern, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *Journal of the American Statistical Association*, **89**, 77-80.

Newey, W.K. and Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **46**, 33-50.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, John Wiley & Sons.

Welsh, A.H. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica*, **6**, 347-366.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Nonparametric Statistics*, **6**, 273-292.