Modelling weather data by approximate regression quantiles.

Hilary M. Green Andrzej S. Kozek*

(Received 1 June 2001; revised 4 June 2002)

Abstract

In the paper we introduce and explore an approximate regression quantiles method. It is based on a new interpretation of M-functionals as quantiles of probability distributions which are determined by the original distribution and the M-function. A correction factor can be applied and this brings the corrected M-functional, called an approximate quantile functional, very close to the quantiles of the original distribution. In the present paper we extend approximate quantile functionals onto parametric models and call them *approximate regression quantiles*. We next model probability distributions of some weather components as they vary over time. We use very simple, but non-linear, parametric models. By applying the approximate regression quantiles method we obtain five-curve summaries of the varying over time probability distributions of the considered weather components.

*Department of Statistics, DEFS, Macquarie University, Australia.
mailto:HGreen@efs.mq.edu.au

[0]See http://anziamj.austms.org.au/V44/CTAC2001/Gree for this article,
 Published 1 April 2003. ISSN 1446-8735

Contents

1 Introduction

Regression quantiles, introduced in [5] and [2], provided a method of estimation of conditional quantiles of parametric regression models. Conditional quantiles in regression models are useful to describe noncentral parts, or even upper and lower boundaries of a cloud of data points. The regression-type data typically has been used to estimate only the behaviour of the central part, that is, the expected value of the conditional distribution also called a regression

curve. The regression quantiles extend this framework onto noncentral parts of the conditional distribution.

In this paper we introduce a method of approximate regression quantiles. The approximate regression quantile method is based on a new simple interpretation of M-functionals as quantiles of probability distributions which are determined by the original distribution and by the M-function (Section 2). Based on this interpretation, a correction factor can be applied (Section 3). This brings the corrected M-functionals, called approximate quantile functionals, very close to conditional quantiles of the original distribution. This correction is next extended onto the case of parametric models, resulting in the approximate regression quantile method discussed in detail in Section 4.

We aim next to model probability distributions of weather components such as temperature, radiation and many other weather factors, as they vary during the day and also on larger, monthly or yearly, time scales. We use very simple, but non-linear, parametric models to describe variability of weather factors over time. By applying the approximate quantiles method we obtain five-curve summaries of the varying over time probability distributions. This results in a simple and comprehensive description of the heteroscedastic nature of the considered data (Section 5).

2 M-functionals and M-estimators

Recall that the expected value can be defined as

$$E\,Y = E_F\,Y = \operatorname*{Arg\,Min}_{\theta} E_F(Y-\theta)^2\,. \tag{1}$$

In a similar way, we have

$$\text{Median}\, Y = \underset{\theta}{\text{Arg Min}}\, E_F|Y-\theta|\,. \tag{2}$$

For general M-functionals, we have

$$Q_M(F) = \underset{\theta}{\text{Arg Min}}\, E_F\, M\,(Y-\theta)\ , \tag{3}$$

where $M(y)$ is a convex function. Hence, the median is a particular M-functional, corresponding to $M(y) = |y|$.

It is important for our method to understand what the M-functional represents. In the remaining part of this section we recall the interpretation of M-functionals given in [6] and extend it to the case of regression quantile modelling.

Let M denote a convex function which has bounded right hand side derivative, $M'(y)$, such that

$$-\infty < -\alpha = \lim_{y\to-\infty} M'(y) < 0 < \lim_{y\to\infty} M'(y) = \beta < \infty\,. \tag{4}$$

Then the derivative of M is bounded and non-decreasing and hence it is a linear function of some cumulative distribution function (CDF), say $G(y)$:

$$M'(y) = (\alpha+\beta)G(y) - \alpha\,. \tag{5}$$

If Y has a CDF F and $p = \beta/(\alpha+\beta)$, then one can show that

$$E_F\, M\,(Y-\theta) - E_F\, M\,(Y) = (\alpha+\beta)\int_0^\theta \Pr\,(Y-Z\le t) - p\,dt\,, \tag{6}$$

where Z is a random variable independent of Y and having G as its CDF. Hence, $Q_M(F)$, *the minimizer of* $E_F\, M\,(Y-\theta)$, *coincides*

with a p-quantile of random variable $U = Y - Z$, *where* $Y \sim F$, Z *is independent of* Y *and* $Z \sim G$.

If $\alpha = \beta$, in particular if function M is symmetric, then the M-functional coincides with the median of $U = Y - Z$. However, if either Y or Z are not symmetric then the medians of Y and U may differ.

Based on (6), we start with a given CDF G and derive the corresponding convex function $M(y)$

$$M_{G,p}(y) = \int_0^y (2G(z) - 1)\, dz + (2p - 1)y\,. \tag{7}$$

Clearly, with $M_{G,p}$ replacing M in (3), we get an M-functional coinciding with a p-quantile of $Y - Z$, $Z \sim G$.

Examples:

1. If
$$M(y) = |y| \tag{8}$$
then $Z = 0$ and
$$G(z) = \begin{cases} 0\,, & \text{if } z \le 0\,, \\ 1\,, & \text{otherwise.} \end{cases} \tag{9}$$
In this case, by using $M_{G,p}(y)$ in (3), we get an M-functional coinciding exactly with the p-quantile of Y. This is a well-known case, cf [7, p.23, Problem 3], being also at the core of regression quantiles introduced in [5].

2. If M is a Huber function (cf [3]) and
$$M_s(y) = \begin{cases} \dfrac{1}{2s}y^2\,, & \text{if } y \in [-s, s]\,, \\ |y| - \dfrac{s}{2}\,, & \text{otherwise}\,, \end{cases} \tag{10}$$

then Z is uniformly distributed on interval $(-s, s)$, that is, $Z \sim U(-s, s)$, and

$$G_s(z) = \begin{cases} 0\,, & \text{if } z \le -s\,, \\ \dfrac{1}{2} + \dfrac{z}{2s} & \text{if } z \in [-s, s]\,, \\ 1\,, & \text{otherwise}\,. \end{cases} \tag{11}$$

By using $M_{G_s,p}(y)$ in (3) we obtain an M-functional coinciding with the p-quantile of $Y - Z$, where $Z \sim U(-s, s)$.

In this paper we shall discuss and use natural estimators of M-functionals, called empirical M-functionals. The empirical M-functionals can be obtained from M-functionals by replacing the CDF F *in* (3) *with the empirical* CDF $\hat{F}_n$.

3 Approximate quantiles

If the convex function is of the form given by formula (7) then we use notation

$$Q_{G,p}(F) = \operatorname*{Arg\,Min}_{\theta} E_F\, M_{G,p}\,(Y - \theta)\ , \tag{12}$$

for the corresponding M-functional defined by (3) with $M = M_{G,p}$.

This functional, $Q_{G,p}(F)$, equals a p-quantile of $U = Y - Z$ and, in general, differs from the p-quantile of F. Therefore we use the following correction:

$$q_p(F) = Q_{G_{1/2}}(F) + \sqrt{\frac{\sigma^2}{\sigma^2 + \mathrm{Var}(G)}}\left(Q_{G_p}(F) - Q_{G_{1/2}}(F)\right)\ , \tag{13}$$

where σ^2 equals the variance of Y and $\mathrm{Var}(G)$ is the variance of random variable Z with CDF G. We call $q_p(F)$ an **approximate quantile functional**.

If $Y \sim N(\mu, \sigma^2)$ and $G(y) = \Phi(y)$, where $\Phi(y)$ is the CDF of the standard normal distribution, then the correction (13) returns exact quantiles of F. In many other cases, in particular for nearly symmetric unimodal distributions with finite variance, the approximation is fairly accurate. The correction works always in the right direction, though, for F very skewed and significantly departing from a normal distribution, it can be less accurate.

4 Approximate regression quantiles and parametric models

Consider now a heteroscedastic regression model

$$Y = g(\mu, x) + \epsilon(x), \tag{14}$$

where x is an explanatory variable, μ is a vector of unknown regression parameters and $\epsilon(x)$ is an error with probability distribution depending on x.

We have seen that $q_p(F)$, given by (13), was picking up a point roughly corresponding to the p-quantile of F. By applying similar method to the parametric model we can fit the parameters of the model to leave approximately $100 \times p$-percent of empirical points below the estimated curve and $100 \times (1-p)$-percent of points above the estimated curve.

Suitable parameters of the model are defined by

$$\mu_p = \operatorname*{Arg\,Min}_{\mu} \int E_F \, M_{G,p}(Y - g(\mu, x)) \, \tau(dx), \tag{15}$$

where τ is a measure of distribution of the design points of the experiment. We call the resulting curve coinciding with the graph of the function

$$y = g\left(\mu_p, x\right) , \tag{16}$$

a regression M-functional of the heteroscedastic distribution of Y. This term is consistent with the *regression quantile* name used in the literature (cf [5]) in the case of $M(y) = |y| + (2p-1)y$. Similarly, as in Section 3 we define approximate regression quantiles, which are based on regression M-functionals and are used in the following part of the paper. The **approximate regression quantiles** are

$$q_p^R(x) = g\left(\mu_{1/2}, x\right) + \sqrt{\frac{\sigma^2}{\sigma^2 + \mathrm{Var}(G)}}(g(\mu_p, x) - g(\mu_{1/2}, x)) , \tag{17}$$

where σ^2 equals the marginal variance of Y and $\mathrm{Var}(G)$ equals the variance of G, with G defined in (5). We shall also consider **empirical regression M-functionals** and **empirical approximate regression quantiles** as estimators of the corresponding population functionals. They are obtained from the population functionals by replacing the population distribution of (x, Y) with the empirical distribution based on $\{(x_1, Y_1), \ldots, (x_n, Y_n)\}$.

We propose to estimate five such approximate regression quantiles curves, corresponding to $p = 0.05$, 0.25, 0.5, 0.75 and 0.95.

Hence, we obtain a five-curve description of the distribution of $Y(x)$. This is analogous to the popular five quantile summary of the Tukey's description of probability distributions of random variables. In the remaining part of this paper we apply the five-curve description of heteroscedastic distributions in the case of various weather components. In all our graphs we use colours: red for $p = 0.05$ and 0.95, blue for $p = 0.25$ and 0.75, and green for $p = 0.5$.

5 Approximate regression quantiles modelling of weather data

We used meteorological data from the Automatic Weather Station, AWS, at Macquarie University (http://atmos.es.mq.edu.au/aws/aws2/) to illustrate applications of approximate regression quantiles.[1]

We used four years of data, from 1994 to 1997, recorded at 15 minute intervals. The data are kept in monthly files.

5.1 Daily and monthly variability

Variability is an important feature of weather data. Models which describe daily variability as well as the general trend of each weather component are advantageous. The crosses in Figure 1 show the global radiation for the first three days (72 hours) in 1995. See that maximum global radiation occurs in the middle of the day. The dots in Figure 2 show the global radiation for the month of January in 1995. The overall daily trend is apparent, as is the heteroscedastic nature of the data.

5.2 Parametric models

For each weather component of interest we proposed a simple model, with as few parameters as possible (maximum of 3), to describe the daily pattern, typical for a given month, as a function of time.

[1] The AWS station is administered by the Atmospheric Science Group, Division of Environmental & Life Sciences, Macquarie University.

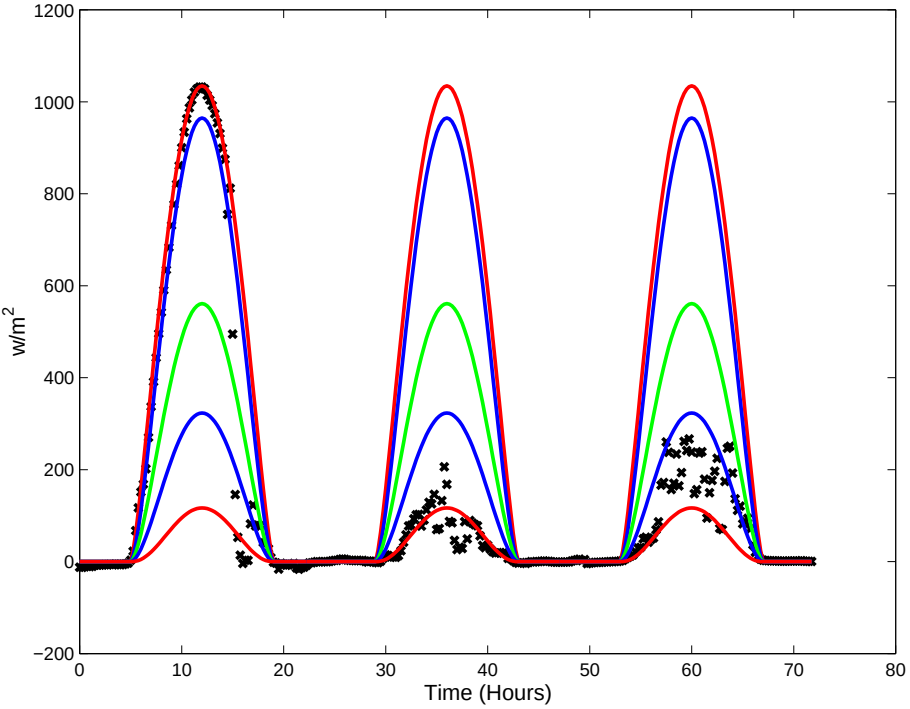

FIGURE 1: Data points and regression quantiles (model (18)) for global radiation over the first 72 hours of January, 1995.

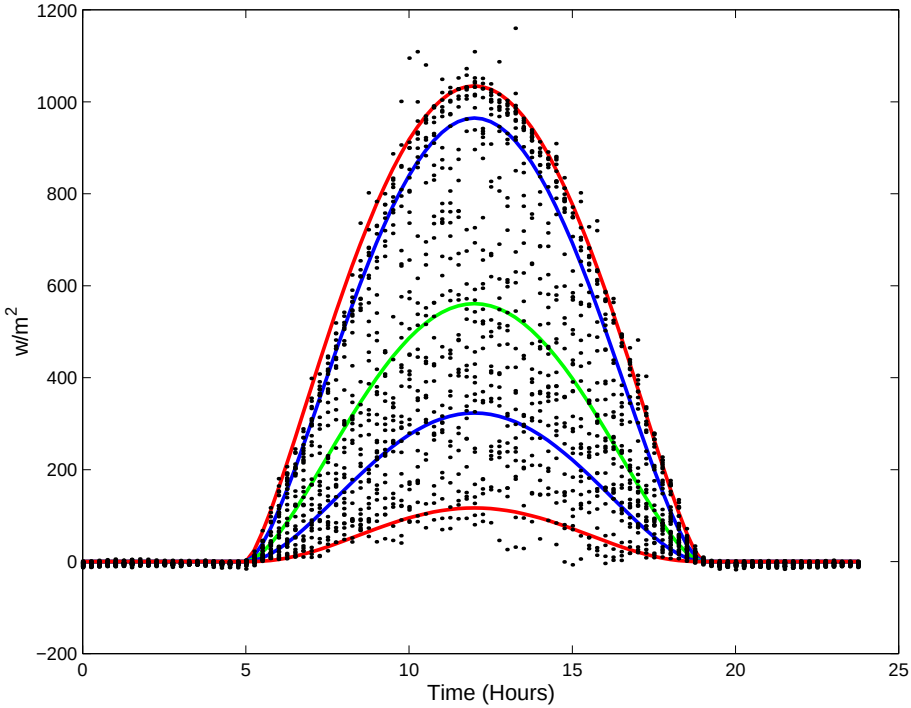

FIGURE 2: Data points and regression quantiles (model (18)) for global radiation in January, 1995, on a daily scale from 0 to 24 hours.

Some simple models we used include: *polynomial functions* to describe radiation data; *cosine functions* to describe wet and dry bulb temperatures, soil temperature, soil heat flux, relative humidity and wind speed; and *maximum of constant and cosine functions* to describe the wind speed data.

Functions of polynomials were chosen to fit the variables when variation was mostly confined to the daylight hours, as in the case of radiation data presented in Figures 1 and 2. The models for these data were of the form

$$Y = \begin{cases} A((X - \mathrm{AM}) \times (\mathrm{PM} - X))^B + \epsilon\,, & \text{for daylight hours}\,, \\ \epsilon\,, & \text{otherwise}\,, \end{cases} \tag{18}$$

where AM and PM are the average monthly times of sunrise and sunset, respectively, A and B are parameters to be estimated, and X is time.

Cosine functions were used where variation occurred over a 24 hour period and in a roughly periodic pattern, for example in the case of temperature data, shown in Figure 3. In general, the minimum daily temperature occurs just before sunrise, and the maximum occurs mid afternoon. The models for these data were of the form

$$Y = A \times \cos\left(\frac{\pi(X + B)}{12}\right) + C + \epsilon\,, \tag{19}$$

where Y is the recorded temperature at time X and A, B and C are the parameters to be estimated.

Maximum of constant and cosine functions model was used to describe the wind speed data shown in Figure 5. We discuss this application in Section 5.5. The models for these data

were of the form

$$Y = \max\left\{C \times \cos(AX + B) + \epsilon, 0.2\right\}\,, \tag{20}$$

where Y is the estimated at time X and A, B and C are parameters to be estimated.

5.3 Numerical methods

To obtain estimates of parameters of regression M-functionals based on models $g\left(\mu_p, x\right)$, discussed above, we solved the minimization problems

$$\hat{\mu}_p = \underset{\mu}{\text{Arg Min}} \frac{1}{n} \sum_{i=1}^{n} M_{G,p}\left(Y_i - g\left(\mu, x_i\right)\right)\,, \tag{21}$$

where we used Huber's M-function (10) and $M_{G,p}$ given by (7). Next, we obtained approximate quantile regressions given by (17). We estimated σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n} \sum r_i^2\,, \tag{22}$$

where

$$r_i = Y_i - g(\hat{\mu}_{1/2}, x_i) \tag{23}$$

were the residuals from fitting the median curve (16) with $p = 1/2$ and $\hat{\mu}_{1/2}$ obtained from (21). All computations leading to estimators of model parameters were carried out in MATLAB.

5.4 Approximate quantile curves for weather data

To describe the data, a simple regression curve is not sufficient because then only the centre of the distribution is estimated and no

description of the heteroscedastic nature of data is reported.

We applied our five-curve technique described earlier to estimate parameters of general curves describing the distribution of the data, conditional on time.

We used the approximate regression quantile method using M-estimators with Huber M-function, with $s = 1$, described in Example 2, to estimate parameters of curves corresponding to the five p-quantiles of the data. In this way we obtained a comprehensive and compact summary of the data. Each curve estimates an approximation to the pth quantile of the distribution corresponding to the model under consideration. For example, approximately 5% of the observed values should be below the fifth quantile curve and 5% should be above the 95th quantile curve.

The lines on Figure 2 show the five estimated regression quantile curves describing, conditional on time, the distribution of global radiation for that month. These regression lines are also shown in Figure 1, for each of the first three days of January, 1995. The first day shows a pattern typical for for the upper 95%, the second day shows a pattern close to the lower 5% and and the third day shows a pattern close to the lower 25% regression curves, respectively.

Similarly, in Figure 3 we show five regression quantile curves for wet bulb temperature. Figure 4 shows how the distribution of global radiation at the middle of the day changes over the four year period, 1994–1997. We plotted the midday values of each of the five regression curves, estimated for all $4 \times 12 = 48$ months. It is interesting that the upper regression curves show clearly the seasonal changes while the lowest 5% curve has a more stable character. The alternating behavior of the 25% and 50% curves is also interesting and not reported earlier. All the distributions are very skewed to the left.

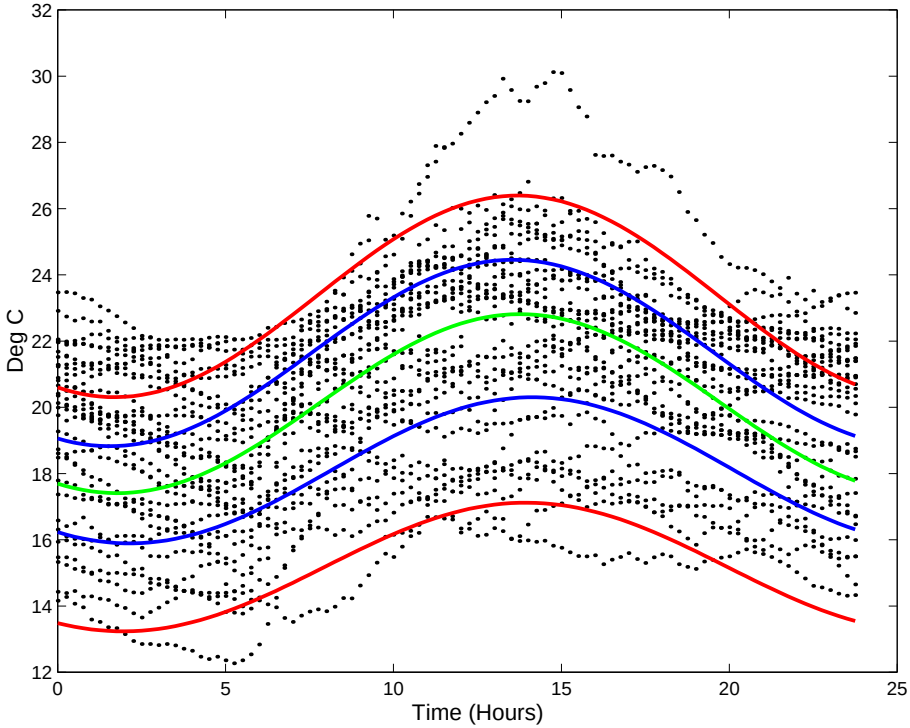

FIGURE 3: Data points and regression quantiles (model (19)) for wet bulb temperature in January, 1995, on a daily scale from 0 to 24 hours.

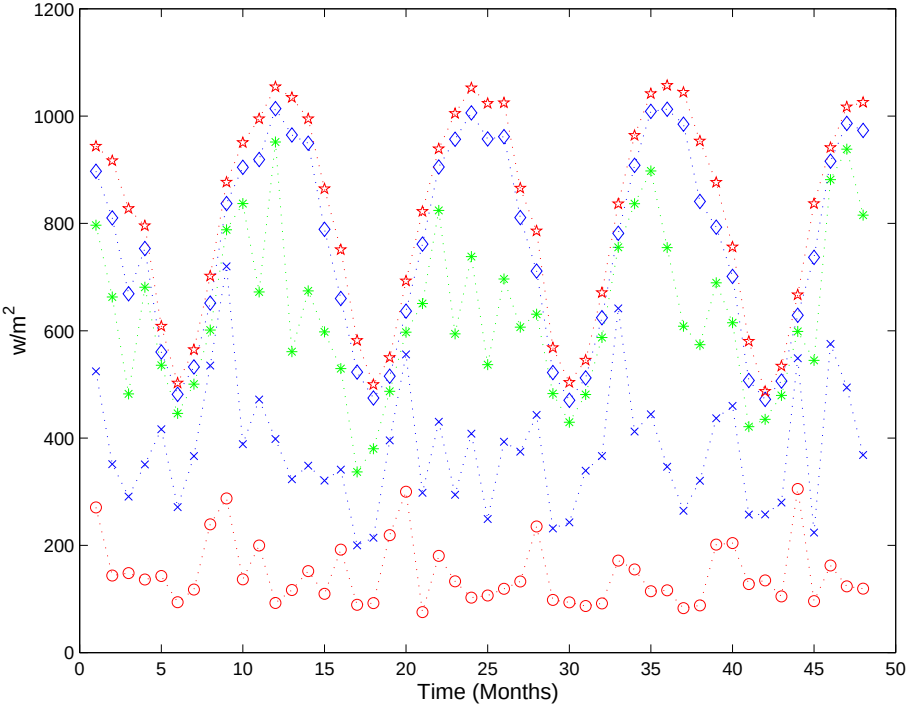

FIGURE 4: Regression quantiles for global radiation at the middle of the day over 48 months: January 1994 – December 1997.

5.5 The correction factor for weather data

Approximate regression quantiles (17) differ from regression M-functionals (16). The approximate regression quantiles are shrunk towards the central, median regression line by the correction factor

$$\kappa = \sqrt{\frac{\sigma^2}{\sigma^2 + \mathrm{Var}(G)}}\,. \tag{24}$$

This correction has descriptive character and tries to compensate for inflation of the distribution of $Y(x)$ caused by using in estimation an M-function corresponding, via (5), to the CDF G. If $\mathrm{Var}(G) \ll \sigma^2$ then $\kappa \approx 1$. This was the case in most of the considered examples, however in some cases, for example, in the case of wind speed data, the factor $\kappa = 0.8325$.

The model considered for the wind speed data was given by (20) and takes into account that the anemometer measuring wind speed records all low wind speeds between 0 and 0.2 m/s as 0.2 m/s. By estimating only regression M-quantile curves (16) we get the top picture in Figure 5. The graphs of the curves are evidently shifted too much apart from the graph of the central, median regression line. For example, there is no data point located below the bottom 5% regression M-quantile.

The approximate regression quantile curves are shrunk towards the central, median regression line, leaving approximately 5% data points below and above of the bottom and top curves, respectively. This example shows that using uncorrected regression M-quantiles may lead to significantly inflated picture of the conditional distribution of $Y(x)$.

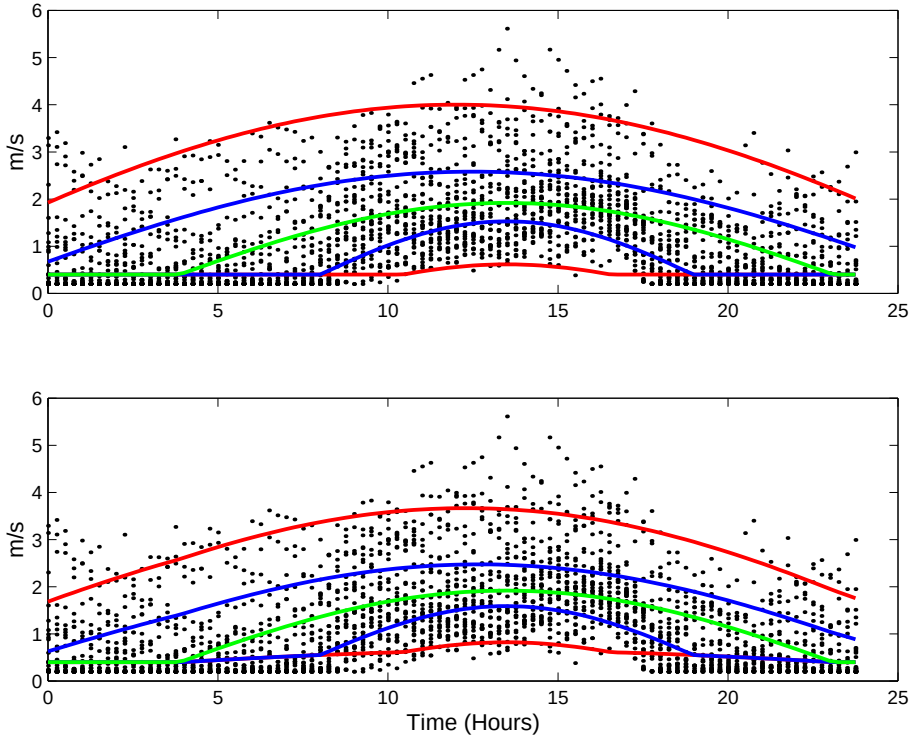

FIGURE 5: Wind speed for August 1994 using model (20), on a daily scale from 0 to 24 hours. Above: regression M-quantiles. Below: approximate regression quantiles.

6 Conclusions

The approximate regression quantiles introduced in this paper are related to the regression quantiles, introduced in [5] and [2] and then extended in a number of papers in various directions, cf [4]. The approximate regression quantiles differ from regression quantiles by using a general convex M-function instead of the absolute value function and by using correction (17).

By using convex functions, different from the absolute value function, and correction (17), one can reduce the variance of the resulting estimators, cf [6]. There is also a numerical advantage. The linear regression quantiles can be computed using a linear programming algorithm [5] or a reduced gradient algorithm for l_1 [8]. In the case of models nonlinear in parameters general minimizing algorithms are used and convex M-functions with continuous derivatives improve their numerical performance.

General convex functions combined with regression models have been considered earlier in the literature, cf [1]. However, by not using correction (17) one may obtain results significantly departing from any sensible modelling of conditional quantiles. Correction (17), based on the novel interpretation of M-functionals is the new contribution of the present paper.

The results obtained by modelling weather components by approximate regression quantiles show that the five curve description of the conditional distribution of weather components is very informative.

References

[1] J. Antoch and P. Janssen. Nonparametric Regression M-Quantiles. *Statistics & Probability Letters*, 8:355–362, 1989. C247

[2] Gilbert Bassett and Roger Koenker. An empirical quantile function for linear models with iid errors. *J. Amer. Statist. Assoc.*, 77(378):407–415, 1982. C230, C247

[3] P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964. C233

[4] Roger Koenker. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *J. Econometrics*, 95(2):347–374, 2000. Principles of econometrics (Madison, WI, 1998). C247

[5] Roger W. Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. C230, C233, C236, C247

[6] A. Kozek. On M-estimators and normal quantiles. *Submitted for publication*, 2001. C232, C247

[7] E. L. Lehmann. *Testing statistical hypotheses.* Wiley, New York, 1991. C233

[8] M. R. Osborne. An effective method for computing regression quantiles. *IMA J. Numer. Anal.*, 12(2):151–166, 1992. C247