The Abdus Salam International Centre for Theoretical Physics

2453-14

School on Modelling Tools and Capacity Building in Climate and Public Health

15 - 26 April 2013

Expectile and Quantile Regression and Other Extensions

KAZEMBE Lawrence

University of Malawi
Chancellor College
Faculty of Science Department of Mathematical Sciences
18 Chirunga Road, 0000 Zomba
MALAWI

Expectile and Quantile Regression and Other Extensions

Lawrence Kazembe
University of Namibia
Windhoek, Namibia

A presentation at
School on Modelling Tools and Capacity Building in Climate and Public Health

ICTP, Trieste, Italy

Objectives and Questions

- Objective:
  Introduce other regression beyond the mean

- Are there median regression models

- What about regression at the mode?

- Can we fit regression model at any other locations too?

- What of the variance, skewness and kurtosis?

Prelimaries

- Given a r.v. $y \sim f(\cdot)$ then
  -$E(Y) = \mu$ defines the mean;
  -$Var(Y) = \sigma^2$ is the variance.

- General assumption: two measures completely define the distribution (**stationarity** concept)

- Classical regression is often characterized by relating to the mean
  -$E(y|\boldsymbol{x}) = \boldsymbol{\beta}x$ if $\boldsymbol{x}$ is the set of covariates.
  -$y \sim N(\mu, \sigma^2)$, then $\mu = \boldsymbol{\beta x}$

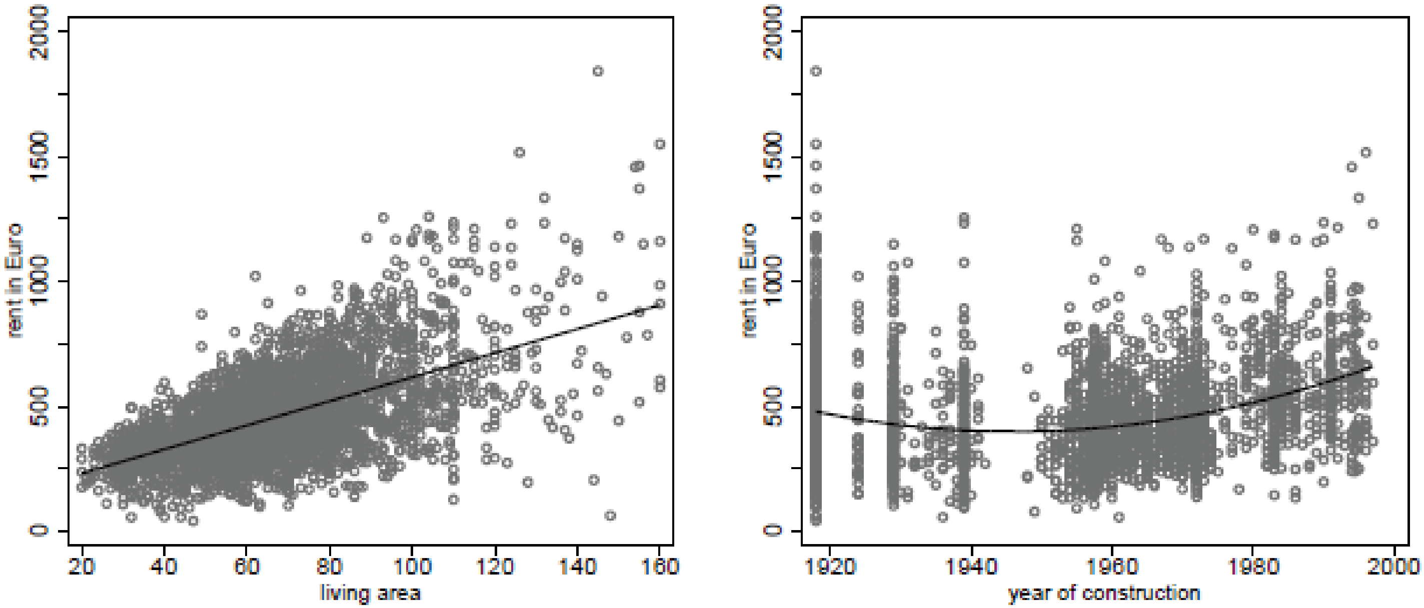
-$y \sim Bern(\pi)$, then $\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{\beta x}$
-$y \sim Pois(\lambda)$, then $\log(\lambda) = \boldsymbol{\beta x}$.

- **Statisticians are mean lovers** (Friedman, Friedman & Amoo)

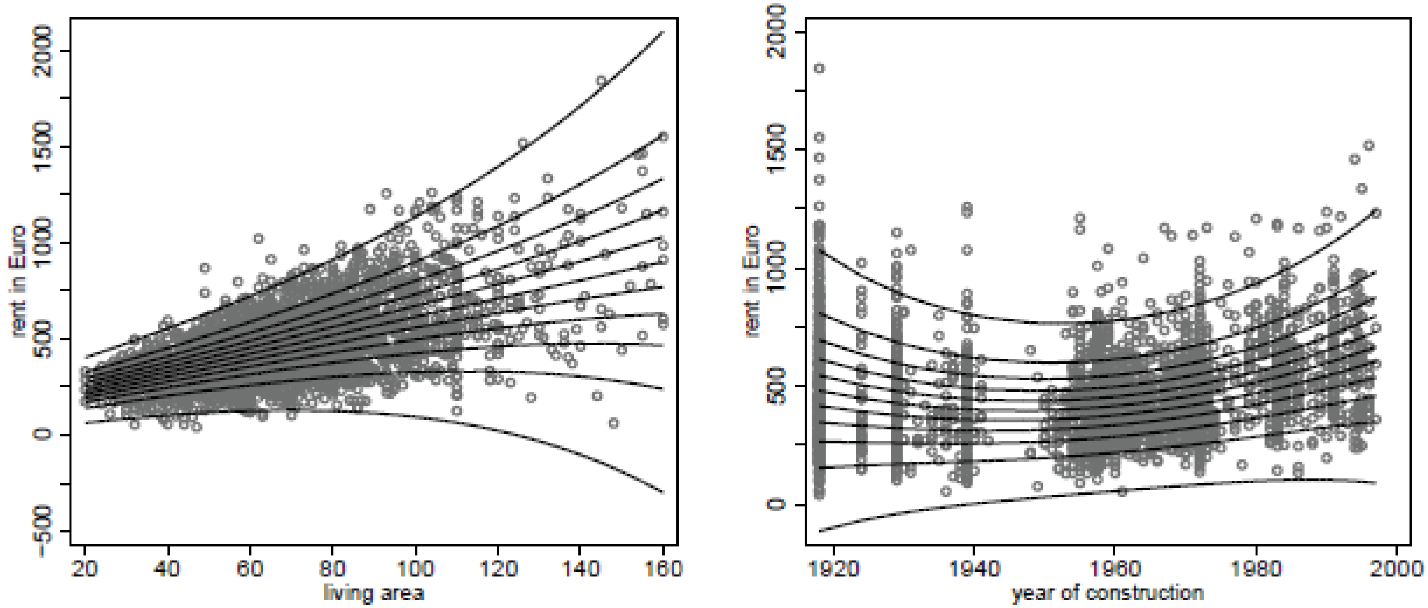
- It goes like:
  We are "mean" lovers. Deviation is considered normal. We are right 95% of the time. Statisticians do it discretely and continuously. We can legally comment on someone's posterior distribution.

- Why the mean?
  -Mean regression reduces complexity
  -However, the mean is not sufficient to describe a distribution.

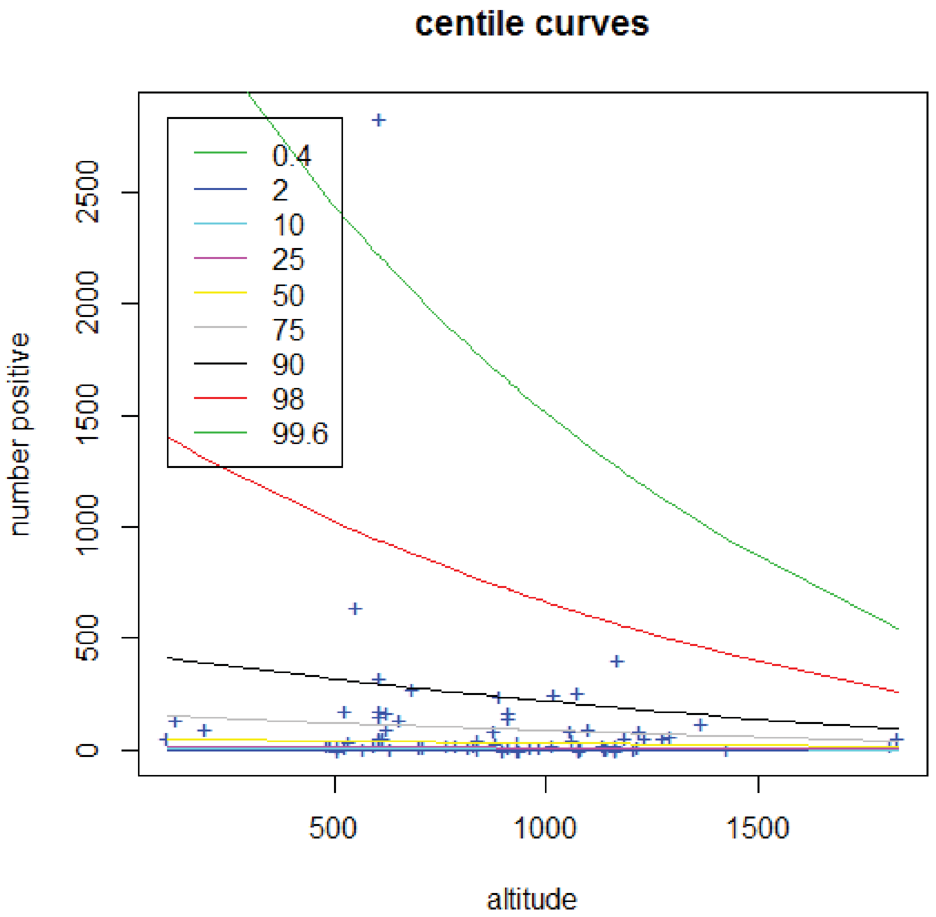
Examples Plot (Rent in Munich, Kneib et al)



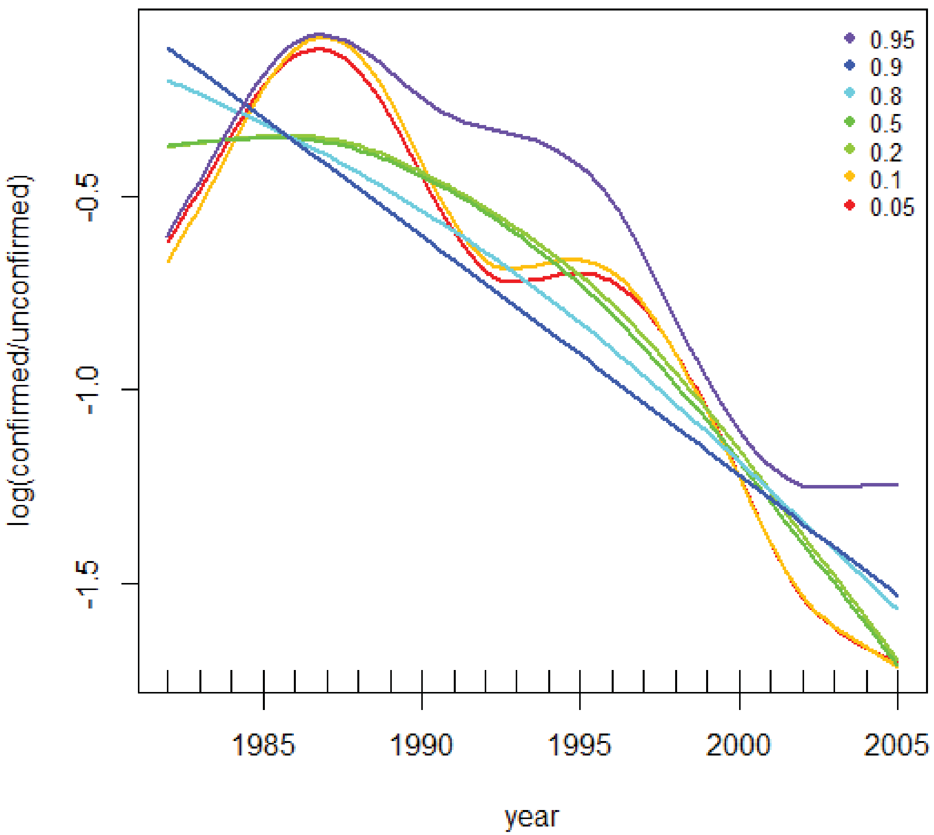
Examples Quantile Plot (Rent in Munich, Kneib et al)



Example (Malaria vs Altitude , Kazembe et al)



Example (Botswana data)



Motivating Examples

- Epidemiology and Public Health:
  -Investigating height, weight and body mass index as a function of different covariates e.g. age (Wei et al. 2006)
  -Exploring stunting curves in African and Indian children (Yue et al. 2012)
  -Generating age-specific centile charts (Chitty et al. 1994).

- Economics:
  -Study of determinants of wages (Koenker, 2005).

- Education:
  -The performance of students in public schools (Koenker and Hallock, 2001).

- Climate data:
  - Spatiotemporal analysis of Boston temperature (Reich 2012)

Double GLM

- Classical regression assumes a homogeneous variance

$$y|\boldsymbol{x}, \varepsilon \sim N(\boldsymbol{\beta x}, \sigma^2)$$

$$\varepsilon \sim N(0, \sigma^2)$$

- However variance heterogeneity (heteroscedasticity) is an order in real-life problems

- The variance of the response may depend on the covariates

- Normal regression example:
  -Regression for mean and variance of a normal distribution

$$y_i = \eta_{i1} + \exp(\eta_{2i})\varepsilon_i, \quad \varepsilon_i \sim N(0,1)$$

such that

$$E(y_i|\boldsymbol{x}_i) = \eta_{1i}$$

$$Var(y_i|\boldsymbol{x}_i) = \exp(\eta_{2i})^2$$

Regression for location, scale and shape

- In general: Specify a distribution for the response on all parameters and relate to the predictors.

- The generalized additive model location, scale and shape (GAMLSS) is a statistical model developed by Rigby and Stasinopoulos (2005).

- For a probability (density) function $f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ conditional on $(\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function to the explanatory variables ($\boldsymbol{X}$)

$$g_1(\mu) = \eta_1 = \boldsymbol{\beta}_1 \boldsymbol{X}_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \quad (1)$$

$$g_2(\sigma) = \eta_2 = \boldsymbol{\beta}_2 \boldsymbol{X}_2 + \sum_{j=1}^{J_2} h_{j2}(x_{j2}) \quad (2)$$

$$g_3(\nu) = \eta_3 = \boldsymbol{\beta}_3 \boldsymbol{X}_3 + \sum_{j=1}^{J_3} h_{j3}(x_{j3}) \quad (3)$$

$$g_4(\tau) = \eta_4 = \boldsymbol{\beta}_4 \boldsymbol{X}_4 + \sum_{j=1}^{J_4} h_{j4}(x_{j4}) \quad (4)$$

- GAMLSS assumes different models for each parameter
  -Model 1: Mean regression
  -Model 2: Dispersion regression
  -Model 3: Skewness regression
  -Model 4: Kurtosis regression

- Other summary measures are also permissible

Quantile Regression

- Quantile, Centile, and Percentile are terms that can be used interchangeably
  -A 0.5 quantile $\equiv$ 50 percentile, which is a median.

- Related terms are quartiles, quintiles and deciles: divides distribution into 4, 5, and 10 parts

- Quantiles are related to the median.

- Suppose Y has a cumulative distribution $F(y) = P(Y \leq \tau)$. Then τth quantile of Y is defined to be

$$Q(\tau) = \int_{-\infty}^{\tau} f(y)dy = \text{inf}\{y : \tau \leq F(y)\}$$

for $0 < \tau < 1$.

- The quantile regression drops the parametric assumption for the error/ response distribution.

- Fit separate models for different asymmetries $\tau \in [0, 1]$:

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau}$$

- Instead of assuming $E(y_i \leq \eta_{i\tau} = 0$, one assumes

$$P(\varepsilon_{i\tau} \leq 0) = \tau$$

or

$$F_{y_i}(\eta_{i\tau}) = P(\varepsilon_{i\tau} \leq 0) = \tau$$

- This gives a set of regression function at any assumed quantile value.

- Assumptions:
  -it is distribution-free since it does not make any specific assumption on the type of errors
  -it does not even require i.i.d errors
  -it allows for heterogeneity.

Expectile Regression

- Expectiles are related to the mean, as are quantiles related to the median.

$\sum_{i=1}^{n} \lvert y_i - \eta_i \rvert \to \min$ median regression	$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau}) \lvert y_i - \eta_{i\tau} \rvert \to \min$ quantile regression
$\sum_{i=1}^{n} \lvert y_i - \eta_i \rvert^2 \to \min$ mean regression	$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau}) \lvert y_i - \eta_{i\tau} \rvert^2 \to \min$ expectile regression

where w_τ is the check function defined by

$$w_\tau = \begin{cases} \tau & \text{if } \; y_i > \mu(\tau) \\ 1-\tau & \text{if } \; y_i \le \mu(\tau) \end{cases}$$

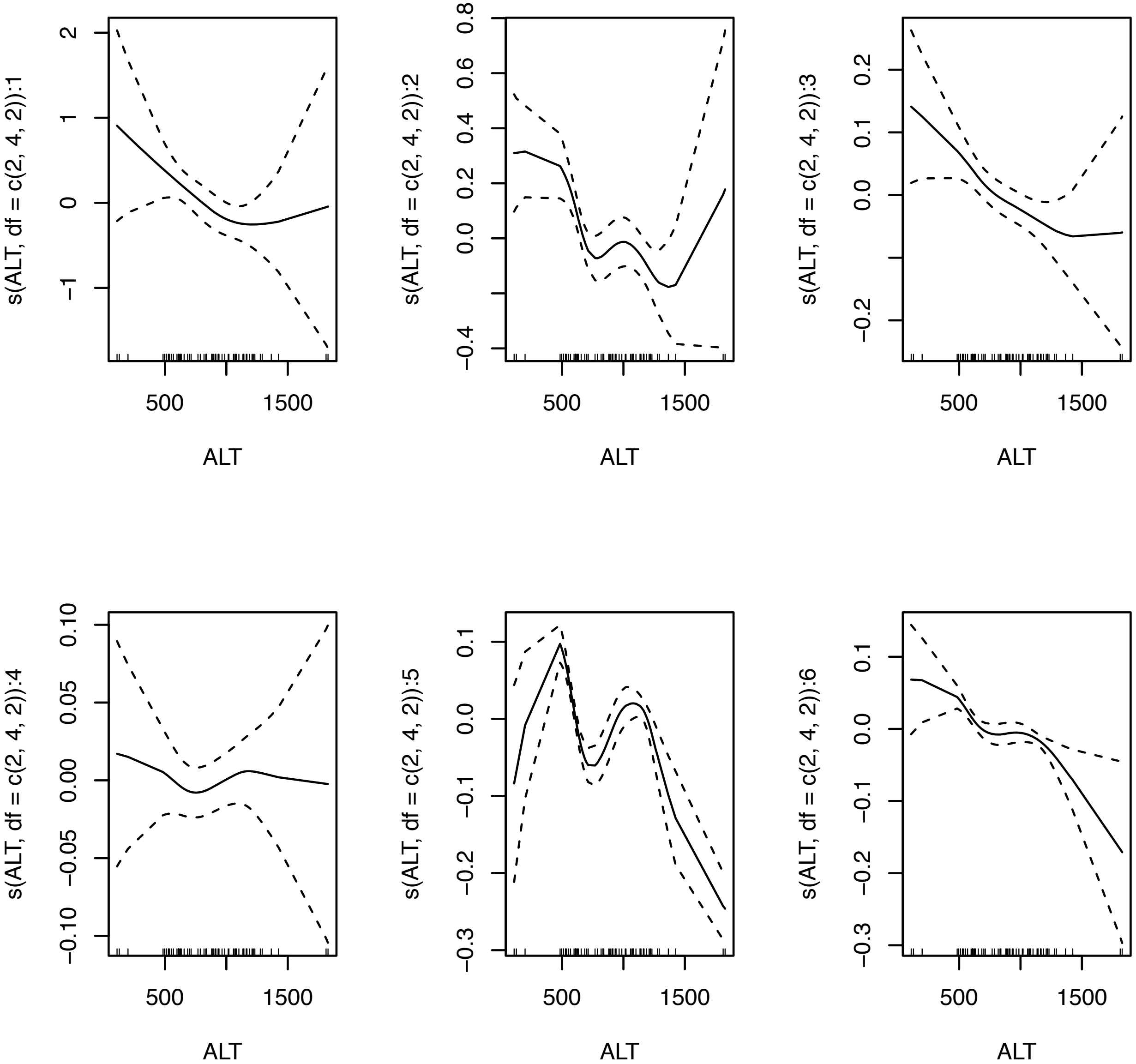
for some population expectile $\mu(\tau)$ for different values of an asymmetric parameter $0 < \tau < 1$.

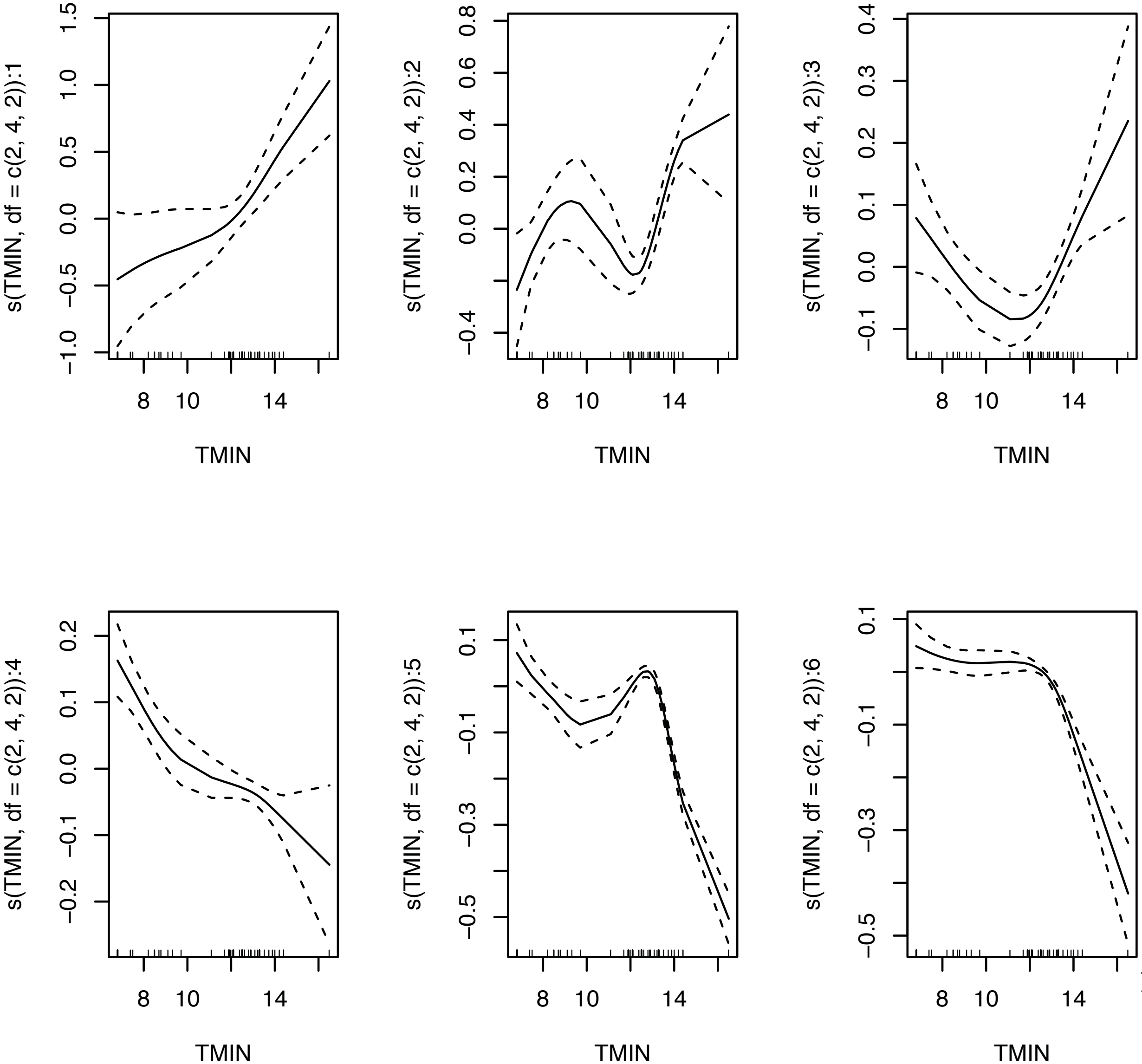
- Expectiles are obtained by solving

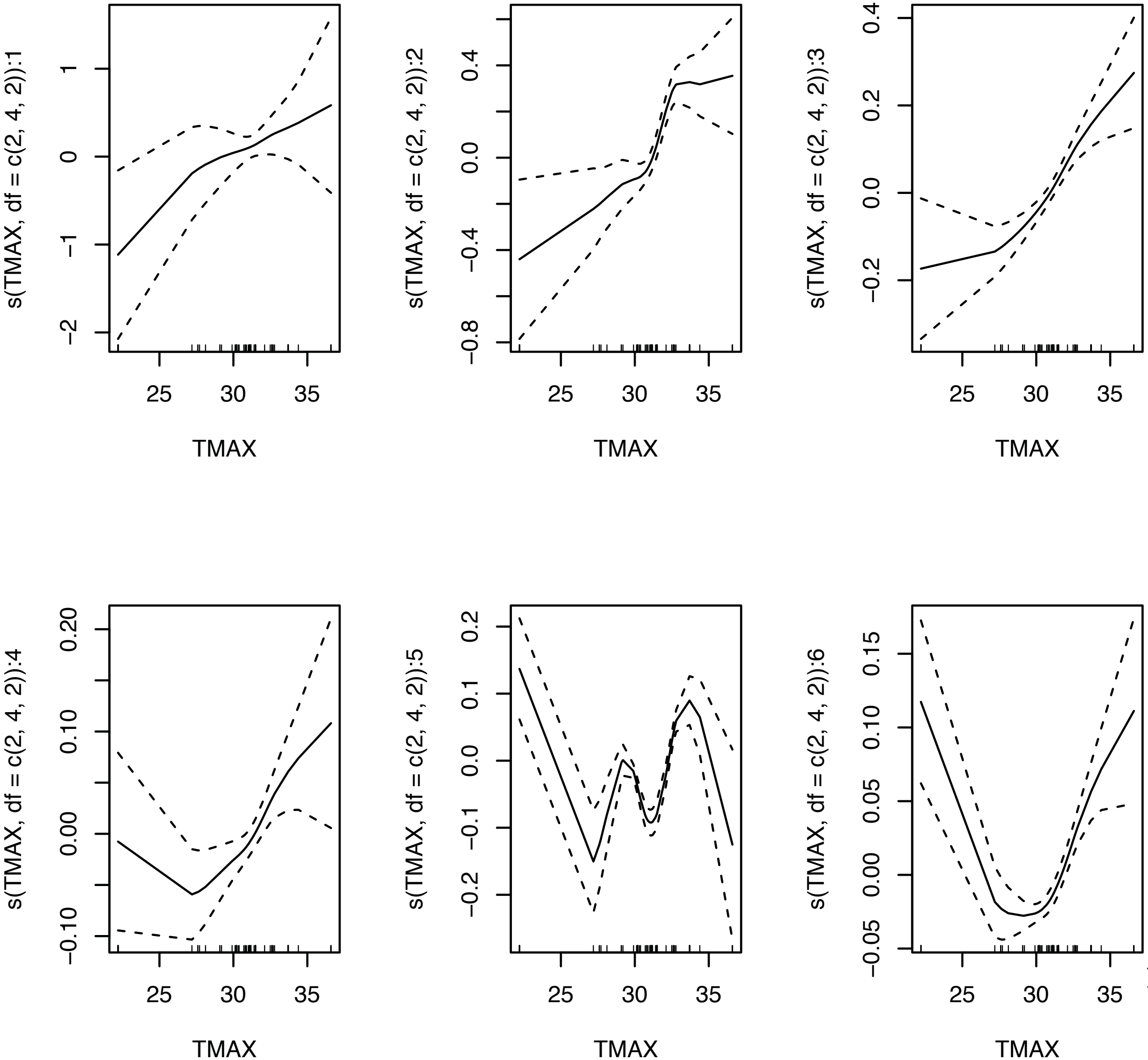
$$\tau = \frac{\int_{-\infty}^{e_\tau} |y - e_\tau| f_y(y) dy}{\int_{-\infty}^{\infty} |y - e_\tau| f_y(y) dy} = \frac{G_y(e_\tau) - e_\tau F_y(e_\tau)}{2(G_y(e_\tau) - e_\tau F_y(e_\tau)) + (e_\tau - \mu)}$$

where
-$f_y(\cdot)$ and $F_y(\cdot)$ denote the density and cumulative distribution function of y.
-$G_y(e) = \int_{-\infty}^{e} y f_y(y) dy$ is the partial moment function of y and
-$G_y(\infty) = \mu$ is the expectation of y.

Worked example - binomial data


s(ALT, df = c(2, 4, 2)):1
2
1
0
−1
500
1500
ALT
s(ALT, df = c(2, 4, 2)):2
0.8
0.6
0.4
0.2
0.0
−0.4
500
1500
ALT
s(ALT, df = c(2, 4, 2)):3
0.2
0.1
0.0
−0.2
500
1500
ALT
s(ALT, df = c(2, 4, 2)):4
0.10
0.05
0.00
−0.05
−0.10
500
1500
ALT
s(ALT, df = c(2, 4, 2)):5
0.1
0.0
−0.1
−0.2
−0.3
500
1500
ALT
s(ALT, df = c(2, 4, 2)):6
0.1
0.0
−0.1
−0.2
−0.3
500
1500
ALT


s(TMIN, df = c(2, 4, 2)):1
s(TMIN, df = c(2, 4, 2)):2
s(TMIN, df = c(2, 4, 2)):3
s(TMIN, df = c(2, 4, 2)):4
s(TMIN, df = c(2, 4, 2)):5
s(TMIN, df = c(2, 4, 2)):6
TMIN
8
10
14


s(TMAX, df = c(2, 4, 2)):1
TMAX
s(TMAX, df = c(2, 4, 2)):2
TMAX
s(TMAX, df = c(2, 4, 2)):3
TMAX
s(TMAX, df = c(2, 4, 2)):4
TMAX
s(TMAX, df = c(2, 4, 2)):5
TMAX
s(TMAX, df = c(2, 4, 2)):6
TMAX

Reference

- Hudson, I. L., Rea, A., Dalrymple, M. L., Eilers, P. H. C (2008). Climate impacts on sudden infant death syndrome: a GAMLSS approach. Proceedings of the 23rd international workshop on statistical modelling pp. 277280.

- Beyerlein, A., Fahrmeir, L., Mansmann, U., Toschke., A. M (2008). Alternative regression models to assess increase in childhood BM. BMC Medical Research Methodology, 8(59).

- Smyth, G. K (1989). Generalized linear models with varying dispersion. J. R. Statist. Soc. B, 51, 4760.

- Sobotka, F., and T. Kneib, 2010. Geoadditive Expectile Regression. Computational Statistics and Data Analysis, doi: 10.1016/j.csda.2010.11.015.