Computational Statistics and Data Analysis 56 (2012) 755–767

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Geoadditive expectile regression

Fabian Sobotka, Thomas Kneib *

*Department of Mathematics, University of Oldenburg, Carl-von-Ossietzky Straße 9–11, 26111 Oldenburg, Germany*

ARTICLE INFO

*Article history:*
Available online 3 December 2010

*Keywords:*
Boosting
Expectiles
Least asymmetric weighted squares
Markov random fields
Quantiles
*P*-splines
Tensor product splines

ABSTRACT

Quantile regression has emerged as one of the standard tools for regression analysis that enables a proper assessment of the complete conditional distribution of responses even in the presence of heteroscedastic errors. Quantile regression estimates are obtained by minimising an asymmetrically weighted sum of absolute deviations from the regression line, a decision theoretic formulation of the estimation problem that avoids a full specification of the error term distribution. Recent advances in mean regression have concentrated on making the regression structure more flexible by including nonlinear effects of continuous covariates, random effects or spatial effects. These extensions often rely on penalised least squares or penalised likelihood estimation with quadratic penalties and may therefore be difficult to combine with the linear programming approaches often considered in quantile regression. As a consequence, geoadditive expectile regression based on minimising an asymmetrically weighted sum of squared residuals is introduced. Different estimation procedures are presented including least asymmetrically weighted squares, boosting and restricted expectile regression. The properties of these procedures are investigated in a simulation study and an analysis on rental fees in Munich is provided where the geoadditive specification allows for an analysis of nonlinear effects of the size of flats or the year of construction and the spatial distribution of rents simultaneously.



## 1. Introduction

To obtain a complete picture of the dependence of a response on covariate information, a single mean regression analysis is often not sufficient. This is most easily seen for heterogeneous data where the impact of covariates on the mean is different from the impact on the variability or more generally the tails of the distribution of the response. As a consequence, quantile regression (Koenker, 2005) is nowadays routinely applied to regression data since quantile regression results for a dense set of quantiles allow for an analysis of the complete conditional distribution of the response. In this paper, we consider expectile regression (Schnabel and Eilers, 2009) as an alternative possibility for characterising the conditional distribution.

Let $Z$ be a continuous random variable with density $f_Z(z)$. Then the $\tau$-quantile $q_\tau$, $\tau \in (0, 1)$, is defined implicitly by the equation

$$\tau = \mathbb{P}(Z \leq q_\tau) = \frac{\int_{-\infty}^{q_\tau} f_Z(z)\mathrm{d}z}{\int_{-\infty}^{\infty} f_Z(z)\mathrm{d}z}$$

but can also be derived from the minimisation problem

$$q_\tau = \arg\min_q \mathbb{E}\left(w_\tau(Z)|Z - q|\right)$$

* Corresponding author.
*E-mail address:* thomas.kneib@uni-oldenburg.de (T. Kneib).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.11.015

with weights

$$w_\tau(Z) = \begin{cases} \tau & \text{if } Z \geq q \\ 1-\tau & \text{if } Z < q. \end{cases}$$

While the implicit definition characterises the quantile as the partial integral of the density, the minimisation problem describes the quantile as the minimiser of a weighted least absolute expectation problem. Expectiles $e_\tau$, $\tau \in (0, 1)$, are now obtained by replacing the partially integrated density with the partial moment function $G$, yielding

$$\tau = \frac{\int_{-\infty}^{e_\tau} |z - e_\tau| f_Z(z)\mathrm{d}z}{\int_{-\infty}^{\infty} |z - e_\tau| f_Z(z)\mathrm{d}z} = \frac{G(e_\tau) - e_\tau F(e_\tau)}{2(G(e_\tau) - e_\tau F(e_\tau)) + (e_\tau - \mu)} \tag{1}$$

where $G(e) = \int_{-\infty}^{e} z f_Z(z)\mathrm{d}z$ and $G(\infty) = \mu$ is the expectation of $Z$. Again, this implicit definition can be re-expressed as a minimisation problem but with absolute deviations replaced by squared deviations, i.e.

$$e_\tau = \arg\min_e \mathbb{E}\left(w_\tau(Z)|Z - e|^2\right).$$

These derivations indicate that expectiles are an alternative possibility for characterising the distribution of a random variable. In fact, the expectile function uniquely determines the distribution of $Z$.

To include expectiles in regression models, we start with the basic regression specification

$$y_i = \eta_{\tau i} + \varepsilon_{\tau i}, \quad i = 1, \ldots, n,$$

with continuous responses $y_i$, independent errors $\varepsilon_{\tau i}$ and a predictor $\eta_{\tau i}$ depending on an asymmetry parameter $\tau \in (0, 1)$ that defines the tail area of the response distribution that will be analysed. Instead of assuming $\mathbb{E}(\varepsilon_{\tau i}) = 0$ as in mean regression or $\mathbb{P}(\varepsilon_{\tau i} \leq 0) = \tau$ as in quantile regression, we then assume

$$0 = \arg\min_e \mathbb{E}\left(w_\tau(\varepsilon_{\tau i})|\varepsilon_{\tau i} - e|^2\right)$$

which ensures that the predictor $\eta_{\tau i}$ equals the $\tau$-expectile of response $y_i$. Estimation of the regression effects in the predictor $\eta_{\tau i}$ is then achieved by minimising the sum of asymmetrically weighted squared deviations

$$\rho(\eta_\tau) = \sum_{i=1}^{n} w_\tau(y_i)(y_i - \eta_{\tau i})^2. \tag{2}$$

Note that the expectile regression specification is semiparametric like usual quantile regression models, since apart from independence of the errors and the condition on the expectiles, no further assumptions on the error terms are included. In particular, errors are allowed to be heteroscedastic and they may follow different types of distribution.

The main advantage of expectiles over quantiles is that the criterion (2) is differentiable with respect to regression effects. This will allow us to derive a simple iteratively weighted least squares procedure for estimating expectile-specific regression coefficients which is of particular value in complex regression specifications including nonlinear, random or spatial effects. These extended modelling components usually rely on quadratic penalties which do not immediately fit into the linear programming framework considered for quantile regression. Moreover, expectiles contain the expectation as a special case (with $\tau = 0.5$) such that mean regression is a special case of expectile regression. This also indicates that expectile regression is closer to the concept of explained variance in least squares estimation and expectile-specific parameters can be interpreted with respect to variance heteroscedasticity.

It has also been claimed that expectiles make more efficient use of the available data as compared to quantiles (Newey and Powell, 1987) since they rely on the distance of observations from the regression predictor while quantiles only use the information on whether an observation is above or below the predictor. Of course, this advantage comes at the price of increased outlier sensitivity. Finally, expectile regression provides a smooth family of functions. While quantile regression lines have to go exactly through $p$ points when $p$ is the number of regression coefficients, expectile regression does not have such a restriction.

In the application on the Munich rental guide that will be described in more detail in Section 4, we aim at analysing the impact of covariates on the net rent per square meter in a geoadditive regression model like

$$\text{rent} = \boldsymbol{x}'\boldsymbol{\beta} + f_1(\text{year}) + f_2(\text{size}) + f_{\text{spat}}(\text{district}) + \varepsilon$$

where $f_1$ and $f_2$ are nonlinear functions of the year of construction and the size of the flat, $f_{\text{spat}}$ is a spatial function defined on the roughly 400 districts within Munich and $\boldsymbol{x}'\boldsymbol{\beta}$ captures further regression effects of categorical covariates describing, for example, special kitchen equipment, the location of the flat in the building, etc. Such geoadditive regression specifications have gained considerable attention in mean regression; see Kamman and Wand (2003) which coined the term geoadditive regression, Fahrmeir et al. (2004) or Ruppert et al. (2003, 2009). Additive models have also been introduced to quantile regression based on for example variational regularisation approaches (Koenker et al., 1994), and boosting for empirical risk minimisation (Fenske et al., 2009) and in a Bayesian framework (Yue and Rue, 2011). Quantile regression has also been