Computer Methods and Programs in Biomedicine 153 (2018) 185–190

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Quantile planes without crossing via nonlinear programming

Alan D. Hutson[1]

Roswell Park Cancer Institute, Department of Biostatistics and Bioinformatics, Elm and Carlton Streets, Buffalo, NY 14623, United States

ARTICLE INFO

Article history:
Received 21 March 2017
Revised 18 September 2017
Accepted 13 October 2017

Keywords:
Bootstrap
Monotonicity
Nonlinear constraints
Computational statistics

ABSTRACT

Background and objective: In this note we propose a nonlinear programming approach for simultaneous fitting of quantile regression models for two or more quantiles. The approach is straightforward, flexible and practical. We apply this approach to a dataset of lactic acid values from a screening dataset in childhood malaria.
Methods: We carry out the fitting of simultaneous quantile regression models using a specific definition of a quantile as an expectation via nonlinear programming methods given certain monotonicity constraints.
Results: We illustrate through simulations and examples that are new approach to simultaneous quantile regression is practical and feasible. The approach is supplemented by providing a bootstrap framework for confidence interval estimation.
Conclusions: Our nonlinear programming approach towards solving the simultaneous quantile regression fitting is shown to be a practical approach that should appeal to statistical practitioners.

1. Introduction

Fitting conditional quantiles as a parametric function of covariates has several applications in the biomedical field from robust regression modeling to modeling processes such as growth of a given biological feature as a function of time. For a specific fixed quantile, e.g. the 0.5th quantile (median), there are several readily available software packages, which are based on more than 3 decades of research; e.g. see Koenker and Ng [6] for a nice background review of the general estimation problem framed in terms of the various linear programming algorithms available for this purpose. It is our goal in this note to extend the ideas of Koenker and Ng [6] to the estimation problem given by fitting several quantile regressions simultaneously, e.g. fitting the 0.25th, 0.5th and 0.75th quantile regression of some outcome of interest as a function of a set of covariates, in the context of a nonlinear programming application. Towards this end we first start with some background notation.

For our purposes, a clean definition of the population quantile function given in terms of an expectation and minimization is as follows:

$$Q_\theta(\alpha) = \inf_{\theta\in\mathcal{R}} E\left\{\frac{|Y-\theta| + (2\alpha-1)(Y-\theta)}{2} - \frac{|Y| + (2\alpha-1)Y}{2}\right\}. \tag{1.1}$$

The properties of (1.1) are outlined in Abdous and Theodorescu [2] who generalize the definition of the α-quantile to $\mathcal{R}^k$ space, $k \geq 1$. Most notably, the assumption that $E(Y) < \infty$, $\alpha \in (0, 1)$, in (1.1) is not a necessary condition for the existence of the α-quantile. The most straightforward case is at $\alpha = 1/2$, which corresponds to the median.

A natural extension to the regression setting is to replace θ in (1.1) with a continuous function $g(\cdot)$ of a set of p covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ (oftentimes setting $x_1 = 1$ when an intercept term is involved) and p regression parameters $\beta' = (\beta_0, \beta_1, \ldots, \beta_{p-1})$. This substitution defines the population conditional quantile function as an expectation and minimization problem of the form

$$Q_\beta(\alpha|\mathbf{x}) = \inf_{\beta\in\mathcal{R}^p} E\left\{\frac{|Y-g(x'\beta)| + (2\alpha-1)(Y-g(x'\beta))}{2} - \frac{|Y| + (2\alpha-1)Y}{2}\right\}. \tag{1.2}$$

For example, if $g(x'\beta) = \beta_0 + \beta_1 x$ and $\alpha = 1/2$, then Eq. (1.2) would correspond to assuming a model for which the conditional median of Y given x is linear in x, i.e. least absolute deviations (LAD) regression or L_1 regression. In the case

E-mail address: ahutson@buffalo.edu

[1] This work was supported by Roswell Park Cancer Institute and National Cancer Institute (NCI) grant P30CA016056, National Science Foundation under grant NSF IIS-1514204 and NRG Oncology Statistical and Data Management Center grant U10CA180822. We wish to acknowledge the 2 reviewers and associate editor for their helpful comments.

https://doi.org/10.1016/j.cmpb.2017.10.019
0169-2607/

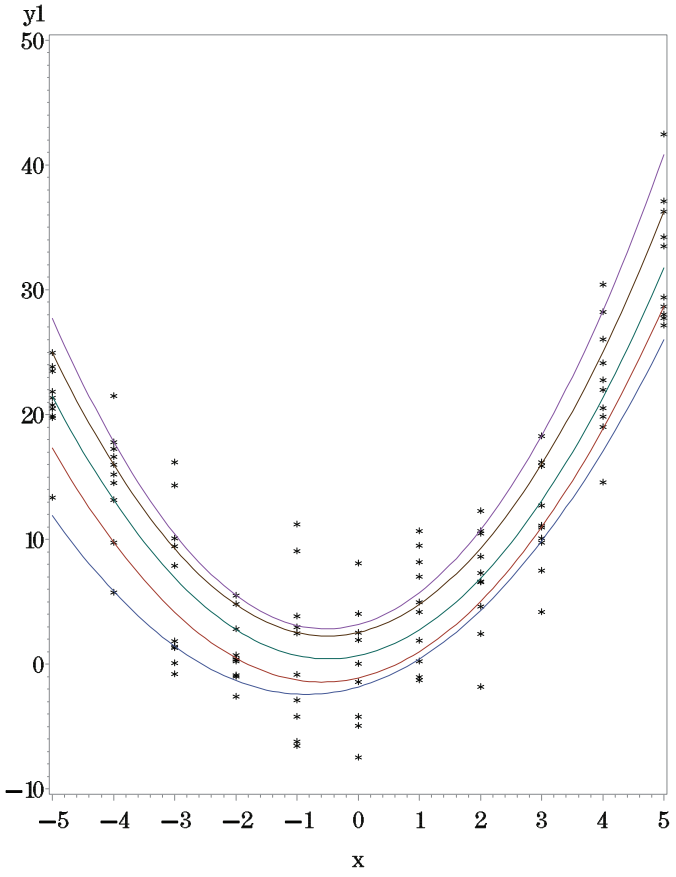

Fig. 1. Model fit for $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon(x)$, where $\epsilon(x) \sim N(0, 5)$, for quantiles $\alpha_1 = 0.1$, $\alpha_2 = 0.25$, $\alpha_3 = 0.5$, $\alpha_4 = 0.75$ and $\alpha_5 = 0.9$.

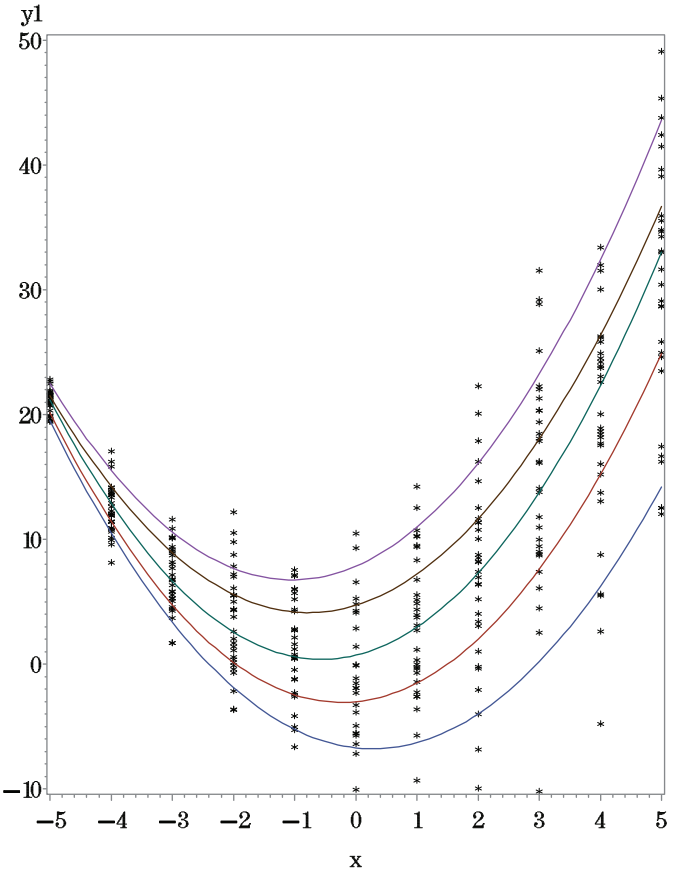

Fig. 2. Model fit for $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon(x)$, where $\epsilon(x) \sim N(0, 6 + x)$, for quantiles $\alpha_1 = 0.1$, $\alpha_2 = 0.25$, $\alpha_3 = 0.5$, $\alpha_4 = 0.75$ and $\alpha_5 = 0.9$.

$\alpha = 1/2$ Eq. (1.2) simplifies to

$$Q_\beta(1/2|\mathbf{x}) = \inf_{\beta \in \mathcal{R}^p} E\left\{ \frac{|Y - g(x'\beta)|}{2} - \frac{|Y|}{2} \right\}. \tag{1.3}$$

In the practical setting, for which we assume Y in (1.2) has no discontinuities, estimation for the αth regression line for a sample size n follows through the minimization of the form

$$\hat{Q}_{\hat\beta}(\alpha|\mathbf{x}) = \inf_{\hat\beta \in \mathcal{R}^p} \sum_{i=1}^{n} |Y_i - g(x_i'\hat\beta)| + (2\alpha - 1)(Y_i - g(x_i'\hat\beta)). \tag{1.4}$$

For example, the estimator for the conditional median regression ($\alpha = 1/2$) then takes the form

$$\hat{Q}_{\hat\beta}(1/2|\mathbf{x}) = \inf_{\hat\beta \in \mathcal{R}^p} \sum_{i=1}^{n} |Y_i - g(x_i'\hat\beta)|. \tag{1.5}$$

The asymptotic theory pertaining to the set of parameter estimates $\hat\beta$ in (1.4) has been well-developed in the classic paper by Koenker and Bassett [7]. In their work they prove that vector $\hat\beta$ has an asymptotic multivariate normal distribution and that $\hat\beta \to \beta$ as $n \to \infty$. There are several software packages that can perform the type of estimation corresponding to Eq. (1.4) for general α, e.g. see SAS's PROC QUANTREG (SAS Institute Inc, Cary, NC, USA) or R's quantreg package [8].

A key issue in estimation in this arena, which is not addressed by the algorithms utilized in the standard software packages, arises when one is interested in simultaneously fitting several quantiles at one time as function of a vector of covariates $\mathbf{x}$, e.g. fitting $Q_{\boldsymbol{\beta}}(\alpha|\mathbf{x})$ for $\alpha = 1/4$, $\alpha = 1/2$ and $\alpha = 3/4$ simultaneously, without even necessarily assuming the same parametric functional form for each respective quantile of interest. This problem has become described throughout the literature in one variant or another as the *quantile crossing problem*. This problem was first discussed roughly 3 decades ago by Bassett and Koenecker [3]. Solutions to specific subsets of this problem have been handled nicely by a variety of authors, but usually only in the two-dimensional single covariate "curve" setting. Chernozhukov et al. [1] provide an outstanding review of these issues to-date. The most notable method for one specific subset of the general problem is the approach of He [4], who develops a modeling approach tied to the assumption of linear heteroscedastic errors, which has a very stringent and limiting assumption in terms of the form of the error term. Chenrozhukov et al. [1] appear to be the first to have a general all-purpose solution to this problem specific to nonparametric estimation of quantile curves. They utilize what they term as a complex sorting algorithm. However, their approach is rather intricate, such that its practical utility for more than a single binary covariate is not obvious. In addition, given the nonparametric nature of their approach, inference about covariates is not directly applicable. More recently Yang and Tokdar [5] introduce a Bayesian approach over arbitrarily shaped convex predictor domains. Their approach is computationally intensive. The authors note that their method requires a complex grid search and can not be applied to moderate to large datasets.

2. Nonlinear constrained quantile regression

Let $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_k < 1$ denote the k quantiles of interest and let $Q_{\beta_1}(\alpha_1|\mathbf{x}), Q_{\beta_2}(\alpha_2|\mathbf{x}), \ldots, Q_{\beta_k}(\alpha_k|\mathbf{x})$ denote the conditional quantiles corresponding to Eq. (1.2), where $\beta_j, j = 1, 2, \ldots, k$, denotes the $p \times 1$ vector of regression parameters corresponding to