Statistics and Probability Letters 117 (2016) 183–191

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Estimation of linear composite quantile regression using EM algorithm

Yuzhu Tian [a,*], Qianqian Zhu [b], Maozai Tian [c]

[a] School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang, China
[b] School of Science, The University of Hong Kong, Hongkong, China
[c] School of Statistics, Renmin University of China, Beijing, China

ARTICLE INFO

Article history:
Received 29 July 2015
Received in revised form 26 May 2016
Accepted 26 May 2016
Available online 1 June 2016

Keywords:
CALD
Composite quantile regression
EM algorithm
AIC (Akaike's information criterion)
BIC (Bayesian information criterion)

ABSTRACT

By incorporating the Expectation–maximization (EM) algorithm into composite asymmetric Laplace distribution (CALD), an iterative weighted least square estimator for the linear composite quantile regression (CQR) models is derived. Two selection methods for the number of composite quantiles via redefined AIC and BIC are developed. Finally, the proposed procedures are illustrated by some simulations.

1. Introduction

Regression models are commonly estimated by traditional least square estimation (LSE). However, LSE is expected to be sensitive to outliers and becomes less efficient when data is non-normal. As an alternative, quantile regression (QR) introduced by Koenker and Bassett (1978) has become an appealing statistical tool in modern regression analysis as well as numerous practical applications such as finance, medicine and environment.

However, Zou and Yuan (2008) showed that QR can lead to an arbitrarily small relative efficiency compared to LSE, and they proposed CQR estimation which combines information of multiple quantiles together to construct a robust and efficient estimation. They indicated that CQR could be much more efficient than LSE. In the context of regression analysis, CQR can provide more robust estimation results for non-normal error distributions than traditional mean regression, even than median regression. Kai et al. (2010) studied local polynomial CQR for nonparametric regression. Kai et al. (2011) conducted CQR estimation and variable selection for semiparametric varying coefficient partially linear models. Jiang et al. (2012) investigated variable selection for nonlinear models via CQR. For more applications of CQR, see Jiang and Qian (2013), Jiang et al. (2014), Jiang and Li (2014), Chen et al. (2015) and Yang et al. (2015). However, it should be mentioned that the objective function to be optimized in CQR is a weighted combination of a sequence of convex functions, which makes the estimation challenging. No closed form of parameter estimators can be derived. Few literatures discussed the optimization for CQR from the viewpoint of computation. Furthermore, dealing with CQR based on likelihood methods is a promising point. To the best of our knowledge, likelihood-based approaches for CQR have not been considered in the literature. Therefore, we propose a likelihood-based method for CQR in this paper. Specially, by introducing the CALD, we study the maximum

* Corresponding author.
E-mail address: pole1999@163.com (Y. Tian).

http://dx.doi.org/10.1016/j.spl.2016.05.019
0167-7152/

likelihood estimation (MLE) for linear CQR models. Based on the EM algorithm, an iterative weighted least square estimator of regression coefficient is obtained. Detailed discussion can be found in Section 3.

The remainder of this paper is organized as follows. In Section 2, we provide preliminary introduction for QR and CQR. In Section 3, we discuss the MLE for linear CQR and employ the EM algorithm to obtain an iterative closed form for parameter estimators. In Section 4, two model selection criteria are used to select a proper number of composite quantiles. In Section 5, simulation studies are conducted to illustrate the finite sample performance of the proposed method. Section 6 draws some conclusions.

2. Preliminary description

2.1. The QR

Consider the common linear regression model

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where y_i is the ith observation, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and ε_i is the error term.

Based on QR in Koenker and Bassett (1978), the τth quantile estimator of regression coefficient β can be obtained as follows

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta), \tag{2}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the quantile check function. From Yu and Moyeed (2001), minimizing the above objective loss function is equivalent to maximizing the likelihood under the asymmetric Laplace distribution (ALD) errors. Probability density function (pdf) of ALD is

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}, \tag{3}$$

where μ is the location, σ is the scale, and $0 < \tau < 1$ is the skewness.

Yu and Moyeed (2001) argued that empirical results are robust by forcing the ALD on errors even if it is a misspecification of the true errors. Additionally, based on the mixture representation of ALD developed by Reed and Yu (2009) and Kozumi and Kobayashi (2011), model (1) can be equivalently rewritten as

$$y_i = x_i^T \beta + \theta_1 \upsilon + \sqrt{\theta_2 \sigma \upsilon_i} \cdot e_i, \quad i = 1, 2, \ldots, n, \tag{4}$$

where $\theta_1 = \frac{1-2\tau}{\tau(1-\tau)}$, $\theta_2 = \frac{2}{\tau(1-\tau)}$, $\upsilon_i \sim \text{Exp}(\frac{1}{\sigma})$, $e_i \sim N(0, 1)$, υ_i and e_i are independent of each other. Representation (4) has been frequently utilized in Bayesian QR papers such as Reich et al. (2011), Kobayashi and Kozumi (2013), Zhao and Lian (2015) and EM algorithm QR papers such as Tian et al. (2014) and Zhou et al. (2014).

2.2. The CQR

Let Q be a set of quantiles of interest, $Q = \{\tau_k, 0 < \tau_1 < \cdots < \tau_K < 1\}$. The CQR estimators of β can be obtained as follows

$$(\hat{\alpha}_1, \ldots, \hat{\alpha}_K, \hat{\beta}^{CQR}) = \arg\min_{\alpha_1, \ldots, \alpha_K, \beta} \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{\tau_k}(y_i - x_i^T \beta - \alpha_k), \tag{5}$$

where α_k's are the τ_kth quantiles of ε_i which satisfy: $\alpha_1 < \cdots < \alpha_K$. To ensure the identifiability of β, we further assume that $\sum_{k=1}^{K} \alpha_k = 0$. Generally, one can set $\tau_k = \frac{k}{K+1}$, $k = 1, \ldots, K$, where K is the number of quantiles level. It should be noted that (5) is a combination of a series of objective functions for QR under several quantiles, and median regression is a special case for $K = 1$.

We mention that objective function (5) is a combination of a sequence of convex functions, which makes the estimation complicated. We cannot obtain closed form of parameter estimators by directly differentiating the objective function. Few papers discussed the optimization for CQR from the viewpoint of computation. Based on Bayesian statistics, Tian et al. (submitted for publication) introduced the CALD to conduct CQR for model (1). However, in Bayesian literatures, statistical inference is heavily dependent on the well-known Bayesian factor. Bayesian factor suffers from various theoretical and computational difficulties. For example, it cannot be defined under improper prior distributions, see, e.g., Li and Yu (2012). In the following sections, we will address CQR problem via the popular EM algorithm.