Computational Statistics and Data Analysis 57 (2013) 558–569

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Assessing model adequacy in possibly misspecified quantile regression

Hohsuk Noh *, Anouar El Ghouch, Ingrid Van Keilegom

Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, 20 Voie du Roman Pays, Louvain-la-Neuve 1348, Belgium

ARTICLE INFO

Article history:
Received 5 January 2012
Received in revised form 20 July 2012
Accepted 21 July 2012
Available online 27 July 2012

Keywords:
Coefficient of determination
Conditional quantiles
Lack-of-fit
Linear model
Prediction quality

ABSTRACT

Possibly misspecified linear quantile regression models are considered. A measure for assessing the combined effect of several covariates on a certain conditional quantile function is proposed. The measure is based on an adaptation to quantile regression of the famous coefficient of determination originally proposed for mean regression, and compares a 'reduced' model to a 'full' model, both of which can be misspecified. An estimator of this measure is proposed and its asymptotic distribution is investigated both in the non-degenerate and the degenerate case. The finite sample performance of the estimator is studied through a number of simulation experiments. The proposed measure is also applied to a data set on body fat measures.

1. Introduction

Quantile regression has emerged as an attractive alternative to the classical mean regression based on the quadratic loss function. Since it was introduced by Koenker and Bassett (1978) as a robust (to outliers) and flexible (about error distribution) linear regression method, quantile regression has received considerable interest in both theoretical and applied statistics (see Koenker, 2005 and references therein). As this method has widened its applications to many domains like economics, biology, ecology and finance, it becomes very attractive for practitioners to develop a simple and effective assessment measure of goodness-of-fit for quantile regression in the spirit of the well-known R^2: the coefficient of determination for linear mean regression.

To this end, when it is assumed that the true conditional quantile function is linear with respect to the given covariates, Koenker and Machado (1999) proposed a measure of model adequacy in quantile regression that aims at measuring prediction quality of a certain set of covariates. However, theoretical properties of its estimator have not been studied except for the work of Mckean and Sievers (1987), who showed the consistency of the estimator for the conditional median. Further, no results exist when the linear model is misspecified. The consideration of misspecification deserves attention because it is rare in reality that such a linear model assumption is correct. Yet practitioners often prefer to use a linear model for reasons of parsimony or interpretability even when they are not sure about the correct specification.

Motivated by these observations, we revisit the measure proposed by Koenker and Machado (1999) and reinterpret it as a way to compare two nested linear quantile models regardless of whether the corresponding true conditional quantile functions are linear or not. This kind of reinterpretation is made possible because of the result in Kim and White (2003) and Angrist et al. (2006), who showed the consistency for certain "pseudo-true" parameter values and the asymptotic normality of the quantile estimator when the model is misspecified. Further, we provide an asymptotic representation of the estimator of the proposed measure, which implies its consistency and asymptotic normality. This result is meaningful in that such a representation is not known (to the best of our knowledge) even for the case of a correctly specified linear model. Although

* Corresponding author. Tel.: +32 10 479403; fax: +32 10 473032.
E-mail address: word5810@gmail.com (H. Noh).
URL: http://perso.uclouvain.be/ingrid.vankeilegom/DataProgramCSDA.zip (I. Van Keilegom).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.07.020

our main goal in this paper is to propose a simple assessment measure of quantile regression under misspecification, it is possible to use the asymptotic results for a specification test as is shown in Section 4.

The rest of this paper is organized as follows. In Section 2 we introduce our measure and provide some insight into it using a few examples. We provide asymptotic results about the estimator of the proposed measure in Section 3 and describe statistical inference based on it in Section 4. In Section 5 we present some Monte Carlo evidence of the developed theory, whereas the analysis of data on body fat measures is given in Section 6. All the theoretical proofs are deferred to the Appendix.

2. Model adequacy measure under misspecification

To introduce our measure, we bring in some notations. Define the check loss function for a fixed quantile level $q \in (0, 1)$ as $\rho_q(u) = 2u(q - I(u < 0))$. Let Y be a one-dimensional dependent variable and $\mathbf{X} = (\mathbf{X}_0^\top, \mathbf{X}_1^\top)^\top$ be a random covariate vector of dimension $d_0 + d_1$, with $d_0, d_1 \geq 1$. The first (and only the first) element of $\mathbf{X}_0$ is 1. Then, our measure $\zeta(q)$ for assessing the effect of $\mathbf{X}_1$ under possible misspecification of a linear q-th quantile regression model is defined as

$$\zeta(q) = 1 - \frac{\mathrm{E}[\rho_q(Y - \mathbf{X}^\top \boldsymbol{\beta}_q^*)]}{\mathrm{E}[\rho_q(Y - \mathbf{X}_0^\top \boldsymbol{\beta}_{0,q}^*)]}, \tag{1}$$

where $\boldsymbol{\beta}_{0,q}^*$ and $\boldsymbol{\beta}_q^*$ are pseudo true parameters in the sense that they are assumed to be the unique minimizers of $\mathrm{E}[\rho_q(Y - \mathbf{X}_0^\top \mathbf{b}_0)]$ and $\mathrm{E}[\rho_q(Y - \mathbf{X}^\top \mathbf{b})]$ with respect to $\mathbf{b}_0$ and $\mathbf{b}$, respectively. Equivalently, we can say that they are the best approximations to the true quantile regression function that can be found within the two given families of linear models. Neither of the two linear models are supposed to be correct, and they are both possibly subject to model misspecification. In terms of the check loss distance, $\mathrm{E}[\rho_q(Y - \mathbf{X}^\top \boldsymbol{\beta}_q^*)]$ represents the amount of variation of Y that cannot be explained through a 'full' but possibly incorrect linear model in $\mathbf{X}$, and $\mathrm{E}[\rho_q(Y - \mathbf{X}_0^\top \boldsymbol{\beta}_{0,q}^*)]$ is the variation of Y that cannot be explained through the reduced linear model, which is also possibly incorrect. Consequently, $\zeta(q)$ is nothing but the relative loss of explained variation in terms of the check distance that can be attributed to the lack-of-fit of the reduced qth quantile linear model compared to the full one. From the definition, it is clear that $0 \leq \zeta(q) \leq 1$. $\zeta(q) = 0$ is equivalent to saying that $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_0^*, \mathbf{0})$, i.e. no information is lost when considering only the restricted linear model. Note that unlike R^2, which is a global measure, several $\zeta(q)$'s (for different values of q) are able to show a more complete picture of the effect of $\mathbf{X}_1$ both in the center and the tails. In other words, the proposed $\zeta(q)$ is a local measure of goodness of fit.

As noted by a referee, the proposed measure $\zeta(q)$ is only informative about the prediction qualities of the full and reduced models for the outcome under the asymmetric loss function, but it is not directly informative about the approximation properties of these models to the conditional quantile function under a precise loss function. To judge the approximation quality one might consider measuring a distance between $Q_q(Y|X)$ and $X^T\beta^*$ and compare it with the distance between $Q_q(Y|X_0)$ and $X_0^T\beta_0^*$. However, such a comparison is problematic because the target functions, $Q_q(Y|X)$ and $Q_q(Y|X_0)$, are different and further we need to use nonparametric methods to estimate those quantities. A possible solution to this problem could be to use partial quantile regression as explained in Angrist et al. (2006) or to consider a different type of coefficient as proposed in Noh et al. (in press).

Before moving into theoretical analysis of $\zeta(q)$ and its estimator, first we will present two interesting applications of $\zeta(q)$ when $d_0 = 1$ (coefficient of determination) and $d_1 = 1$ (covariate importance). We use the following model for illustration of $\zeta(q)$:

$$Y_i = 0.5 + X_{1i} - 2X_{2i} + \beta_3 X_{3i} + \nu g(X_{1i}) + \sigma\varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $g(X_{1i}) = 1 - \cos(\pi X_{1i}/2)$, $\sigma = 0.4$ and the $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})^\top$'s are i.i.d. and generated from a truncated multivariate normal distribution with the constraints $0 \leq X_{ki} \leq 1$, $k = 1, 2, 3$. The errors ε_i's are independent standard normal variables and are independent of the $\mathbf{X}_i$'s. The details of this model are given in Example 3 of Section 5. Note that ν controls the degree of the linear quantile regression model. Because it is difficult to calculate the value of $\boldsymbol{\beta}_q^*$ and $\boldsymbol{\beta}_{0,q}^*$ analytically, we use a Monte Carlo approximation for the computation of $\zeta(q)$, using a large sample of size 500,000.

2.1. Coefficient of determination ($d_0 = 1$)

If $d_0 = 1$, then $\boldsymbol{\beta}_{0,q}^*$ becomes $\xi_q = \arg\min_b \mathrm{E}[\rho_q(Y - b)]$, which is the marginal qth quantile of Y, and then $\zeta(q)$ becomes

$$\zeta(q) = 1 - \frac{\mathrm{E}[\rho_q(Y - \mathbf{X}^\top \boldsymbol{\beta}_q^*)]}{\mathrm{E}[\rho_q(Y - \xi_q)]} \tag{3}$$

which we will call $R^*(q)$ hereafter. $R^*(q)$ is the quantile analogue of the well known Pearson's correlation ratio $\eta^2 = 1 - \mathrm{E}[Y - \mathbf{X}^\top \boldsymbol{\beta}^*]^2 / \mathrm{E}[Y - \mathrm{E}(Y)]^2$, i.e. the 'theoretical' R^2 for the linear mean regression model $\mathrm{E}(Y|\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}^*$. For this reason, when the underlying distribution is asymmetric or in the presence of outliers, $R^*(0.5)$ could be used as a robust alternative to R^2. Like η^2, $R^*(q)$ lies in $[0, 1]$. $R^*(q) = 0$ corresponds to the case when $\mathbf{X}^\top \boldsymbol{\beta}_q^* = \xi_q$ with probability one, i.e. all components of $\boldsymbol{\beta}_q^*$ vanish except the first one, which coincides with ξ_q. In that case, no variability is captured by $\mathbf{X}$ via a linear