A new lack-of-fit test for quantile regression models using logistic regression

Mercedes Conde-Amboage[1] & Valentin Patilea[2] & César Sánchez-Sellero[1]

[1] *Department of Statistics and O.R.,University of Santiago de Compostela, Spain*
email: mercedes.amboage@usc.es, cesar.sanchez@usc.es
[2] *Center of Research in Economics and Statistics (Ensai), France*
email: patilea@ensai.fr.

Résumé. On propose un nouvel test d'adéquation pour les modèles de régression quantile. Le test se base sur l'interprétation des résidus de la régression quantile comme des variables à expliquer dans une régression logistique associée, avec des variables explicatives bien choisies. La validité du modèle de régression quantile implique la nullité de tous les coefficients dans le modèle logistique associé. L'idée est alors d'utiliser un test de rapport de maximum de vraisemblance dans le modèle logistique pour valider le modèle de régression quantile initial. La nouvelle approche pour vérifier l'adéquation d'une régression quantile détecte des alternatives générales. Une procédure de réduction de la dimension dans la régression logistique associée, à l'aide de projections, est également proposée. Les valeurs critiques pour la statistique de type rapport de vraisemblance sont calculées à l'aide d'une procédure de bootstrap dans le modèle quantile, similaire à celle proposée par Feng et al. (2011). Les simulations montrent que le nouveau test se comporte mieux que les tests non paramétriques existants.

Mots-clés. régression quantile, test d'adéquation, régression logistique, bootstrap.

Abstract. A new lack-of-fit test for parametric quantile regression models is proposed. The test is based on interpreting the residuals from the quantile regression model fit as response values of a logistic regression, the predictors of the logistic regression being functions of the covariates of the quantile model. Then a correct quantile model implies the nullity of all the coefficients but the constant in the logistic model. Given this property, we use a likelihood ratio test in the logistic regression to check the quantile regression model. In the case of multivariate quantile regressions, to avoid working in very large dimension logistic regression, we use predictors obtained as functions of univariate projections of the covariates from the quantile model. Finally, we look for a 'least favorable' projection for the null hypothesis of the likelihood ratio test. Our test can detect general departures from the parametric quantile model. To approximate the critical values of the test, a wild bootstrap mechanism is used, similar to that proposed by Feng et al. (2011). A simulation study shows the good properties of the new test versus other nonparametric tests available in the literature.

Keywords. quantile regression, lack-of-fit test, logistic regression, bootstrap.

1 Introduction

Let us consider a quantile regression model denoted by

$$Y = q_\tau(X) + \varepsilon,$$

where $q_\tau(\cdot)$ represents the regression function and the error ε has a conditional τ-quantile equal to zero, that is $\mathbb{P}(\varepsilon \leq 0 | X = x) = \tau$ for almost all x. Koenker & Bassett (1978) proposed a linear quantile modeling. Chaudhuri (1991) studied the nonparametric quantile regression. More recently, semiparametric single-index models were proposed by Kong & Xia (2012). In this paper, we propose a new test of a parametric (linear or nonlinear) quantile regression model against general alternatives.

Consider a sample of independent observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ of the response variable Y and the covariate $X = (X^{(1)}, \cdots, X^{(q)}) \in \mathbb{R}^q$. The covariate vector could have continuous and discrete components. Formally, we address the problem of testing a parametric model of quantile regression

$$H_0 : q_\tau(\cdot) \in \mathcal{M} = \left\{ q_\tau(\cdot, \theta) \ : \ \theta \in \Theta \subset \mathbb{R}^d \right\}, \tag{1}$$

versus a nonparametric alternative $\sup_\theta \mathbb{P}(q_\tau(X) = q_\tau(X, \theta)) < 1$. This problem was addressed by He & Zhu (2003), Zheng (1998), Horowitz & Spokoiny (2002), among others.

For the parameter estimation in the model under test, we follow Koenker & Bassett (1978) who proposed estimating the τ-quantile of the response variable Y given the covariate X (denoted by $q_\tau(X, \theta)$) as the minimizer of

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - q_\tau(X_i, \theta) \right),$$

where $\rho_\tau(u) = u\big(\tau - \mathbb{I}(u < 0)\big)$ is the well-known quantile loss function and $\mathbb{I}(\cdot)$ denotes the indicator of an event. In the following, $\widehat{\theta}$ is a solution of this minimization problem.

2 The new method

Following an idea introduced by Redden et al. (2004), the new lack-of-fit test is based on the dichotomous variable

$$Z(\theta) = \mathbb{I}(Y \leq q_\tau(X, \theta)).$$

The parametric quantile regression model is correct if and only if there exists some $\theta_0 \in \Theta$ such that the conditional probability of $Z(\theta_0)$ given X does not depend on X, and is equal to τ. To check the independence between a suitable $Z(\theta)$ and X, the idea is to consider a logistic regression with response $Z(\widehat{\theta})$ and many covariates obtained as functions of the components of the vector X, and to test the nullity of all the coefficient but the constant.

Let us introduce some more notation. Consider a sample $(W_1, V_1), \cdots, (W_n, V_n)$ where the response variable V only takes values 0 or 1, and W is a vector of explanatory variables with the first component equal to 1. In a logistic regression,

$$\text{logit}\left(\mathbb{P}(V = 1 | W = w)\right) = \varphi' w,$$

where $\text{logit}(p) = \log(p/(1-p))$ is the well-known logistic transformation and φ is a vector of parameters. We can estimate the previous logistic regression model via penalized maximum likelihood (ML). Then, the estimated parameter $\widehat{\varphi}$ could be computed as follows:

$$\begin{aligned}\widehat{\varphi} &= \arg\max_{\varphi} \left[n^{-1} L_n(\varphi, V, W) + \lambda \|\varphi\|\right] \\ &= \arg\max_{\varphi} \left[\frac{1}{n} \sum_{i=1}^{n} \left(V_i \varphi' W_i - \log(1 + e^{\varphi' W_i})\right) + \lambda \|\varphi\|\right],\end{aligned} \tag{2}$$

where L denotes the likelihood function, $\|\cdot\|$ denotes the Euclidean norm and λ is the smoothing parameter. We have considered a penalized ML estimation in order to control for large values of the coefficients that are likely to occur due to the separation problem, a well-known practical aspect in logistic regression.

To detect general alternatives, the vector W used in the logistic regression should contain many functions of the components of X, the original covariate vector in the quantile regression. To avoid working with very large dimensions for W, that are inevitable if X is multivariate, we follow a projection approach. More precisely, we note that H_0 defined in (1) holds true if and only if, for some $\theta_0 \in \Theta \subset \mathbb{R}^d$, and $\forall \beta \in \mathbb{R}^q$ with $\|\beta\| = 1$,

$$\mathbb{E}\left[\mathbb{I}(Y \leq q_\tau(X, \theta_0)) - \tau | F_\beta(\beta' X)\right] = 0, \tag{3}$$

where $F_\beta(t) = \mathbb{P}(\beta' X \leq t)$ represents the distribution function of the projected covariates. This property suggests that it suffices to consider the logistic regression with W a vector of univariate functions of $\beta' X$ and to check whether all the coefficients but the constant are null. Finally, it remains to search a 'least favorable' direction β for the null hypothesis (1), such as Conde-Amboage et al. (2015) and Patilea et al. (2015) did.

To formally describe our procedure, let

$$P_i(\beta) = (1, H_2(\beta' X_i), H_3(\beta' X_i), \ldots, H_p(\beta' X_i))', \qquad 1 \leq i \leq n,$$

represent a basis of Hermite polynomial evaluated at the projections $\beta' X_1, \cdots, \beta' X_n$. Let us recall that the Hermite polynomial of order p is defined as

$$H_p(x) = p! \sum_{m=0}^{[p/2]} \frac{(-1)^m}{m!(p-2m)!} \frac{x^{p-2m}}{2^m}, \qquad x \in \mathbb{R}, \qquad p \geq 1,$$

where $[a]$ denotes the integer part of a real number a. The idea is to check whether, for some value θ_0, we have $\varphi_1 = \varphi_2 = \cdots = \varphi_p = 0$ in logistic regression model

$$\text{logit}\left(\mathbb{P}[Z(\theta_0) = 1 | P(\beta)]\right) = \varphi_1 + \varphi_2 H_2(\beta' X) + \cdots + \varphi_p H_p(\beta' X) = \varphi' P(\beta).$$

The infeasible responses $Z_i(\theta_0)$ are replaced by $Z_i(\widehat{\theta})$. Meanwhile, φ is estimated following the procedure described in equation (2) with

$$L_n(\varphi, Z(\widehat{\theta}), P(\beta)) = \sum_{i=1}^{n} \left(Z_i(\widehat{\theta}) \varphi' P_i(\beta) - \log(1 + e^{\varphi' P_i(\beta)})\right),$$

and some suitable λ. To check the significance of the coefficients φ but φ_1 (the constant φ_1 should be equal to $\text{logit}(\tau)$), we use a likelihood ratio type statistic.

Gathering facts, the new lack-of-fit test for quantile regression is based on the test statistic

$$T = \max_{\beta \in \mathbb{R}^q, \|\beta\|=1} 2\left(L_n(\widehat{\varphi}, Z(\widehat{\theta}), P(\beta)) - L_n(\text{logit}(\tau), Z(\widehat{\theta}), 1)\right), \tag{4}$$

where

$$L_n(\text{logit}(\tau), Z(\widehat{\theta}), 1) = \sum_{i=1}^{n} \left[Z_i(\widehat{\theta})\text{logit}(\tau) - \log(1 + e^{\text{logit}(\tau)}) \right].$$

A bootstrap procedure in the quantile regression context will be proposed in order to calibrate the critical values for the test statistic (4). The bootstrap procedure works as follows:

1.- Let $\varepsilon_i^\star = \delta_i |r_i|$, where $r_i = Y_i - q_\tau(X_i, \widehat{\theta})$ are the residuals from the original sample. The multipliers, δ_i, are independently generated from the two-point distribution with probabilities $(1-\tau)$ and τ at $2(1-\tau)$ and -2τ, respectively; see also Feng et al. (2011). Compute $Y_i^\star = q_\tau(X_i, \widehat{\theta}) + \varepsilon_i^\star$ for each $i = 1, \cdots, n$.

2.- Use the bootstrap data set $\{(X_i, Y_i^\star),\ i = 1, \cdots, n\}$ to compute the bootstrap estimator $\widehat{\theta}^\star$ and the dichotomous variables $Z_i(\widehat{\theta}^\star) = \mathbb{I}(Y_i^\star \leq q_\tau(X_i, \widehat{\theta}^\star))$.

3.- Use the data set $\{(P_i(\beta), Z_i(\theta^\star)),\ i = 1, \cdots, n\}$ to compute the estimator $\widehat{\varphi}^\star$, following the procedure described in (2) with $L_n(\varphi, Z_i(\widehat{\theta}^\star), P(\beta))$, and the new test statistic

$$T^\star = \max_{\beta \in \mathbb{R}^q, \|\beta\|=1} 2\left(L_n(\widehat{\varphi}^\star, Z(\widehat{\theta}^\star), P(\beta)) - L_n(\text{logit}(\tau), Z(\widehat{\theta}^\star), 1)\right).$$

4.- Repeat Steps 1, 2 and 3 many times, and estimate the α-level critical value by the $(1-\alpha)$-quantile of the empirical distribution of $T^\star$.

Note that the covariate of the logistic model, the $P_i(\beta)$ do not need to be computed for each bootstrap sample because it only depends on the covariates. Moreover, to compute the test statistic represented in (4), we are going to use the sequential algorithm based on successive one-dimensional optimizations proposed by Patilea et al. (2015).

3 Simulation study

We will study the performance of the proposed method under the null and alternative hypotheses using Monte Carlo simulations. The number of simulated original samples was 200 and the number of bootstrap replications 500. The number p of Hermite polynomials was $p = [\sqrt{n}]$. In addition, the smoothing parameter λ related to the penalized maximum likelihood will be set to $n^{-1}\log(n)$ in order to avoid the separation problem.

First we study the behavior under the null hypothesis. Data will be simulated from the following median ($\tau = 0.5$) regression model:

$$\textbf{Model 1:} \qquad Y = 1 + X^{(1)} + X^{(2)} + \varepsilon_1$$

	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 49$	0.105	0.035	0.000
$n = 100$	0.100	0.055	0.010

Table 1: Proportions of rejections associated with the new lack-of-fit test for Model 1 for different nominal levels and different sample sizes.

		$\alpha = 0.10$				$\alpha = 0.05$				$\alpha = 0.01$			
		NT	HZ	CSG	MLP	NT	HZ	CSG	MLP	NT	HZ	CSG	MLP
D1	$n = 49$	0.800	0.150	0.130	0.120	0.730	0.090	0.050	0.080	0.555	0.015	0.025	0.025
	$n = 100$	0.985	0.105	0.170	0.155	0.975	0.055	0.085	0.055	0.925	0.015	0.005	0.010
D2	$n = 49$	0.775	0.260	0.075	0.145	0.695	0.160	0.035	0.075	0.485	0.070	0.010	0.025
	$n = 100$	0.995	0.740	0.185	0.145	0.980	0.570	0.055	0.075	0.940	0.220	0.010	0.010

Table 2: Proportions of rejections for each of the three lack-of-fit tests considered for Model 2 for different deviations from the null hypothesis, different sample sizes n and nominal levels α.

where $X^{(i)} \sim U(0,1)$ with $i = 1, 2$ and $\varepsilon_1 \sim N(0,1)$. Table 1 present the proportion of samples for which the null hypothesis (1) was rejected, for different sample sizes n and nominal levels α. The new method shows a good adjustment to the nominal level, even for a small sample size.

Next, the performance of the new test under different alternatives will be studied. The new test will be compared with that of Maistre el at. (2014) denoted by *MLP*, Conde-Amboage et al. (2015) denoted by *CSG*, and He & Zhu (2003) denoted by *HZ*. Our new test will be denoted by *NT*. In order to analyze their performance under the alternative hypothesis, we will consider the following median regression model:

$$\textbf{Model 2:} \qquad Y = 1 + X^{(1)} + X^{(2)} + h(X^{(1)}, X^{(2)}) + \varepsilon_2,$$

where $X^{(1)} \sim N(0,1)$, $X^{(2)} \sim U(0,1)$ and $\varepsilon_2 + 1 \sim LogN(0,1)$. The function $h(\cdot)$ represents the deviation from the null hypothesis. Two different deviations will be considered:

- $h(X^{(1)}, X^{(2)}) = 5\ \sin\left(2\pi(1 + X^{(1)} + X^{(2)})\right)$, that will be denoted by D1;
- $h(X^{(1)}, X^{(2)}) = 10\ (X^{(2)})^2$, that will be denoted by D2.

Table 2 shows the proportion of samples for which the null hypothesis (1) was rejected under each of the cited methods, NT, HZ, CSG and MLP, for different sample sizes n and nominal levels α. According to Table 2, the power of the new test for the deviations D1 and D2 is clearly superior compared with the considered nonparametric competitors.

4 Conclusions

We proposed a new lack-of-fit test for quantile regression models, together with a bootstrap mechanism to approximate the critical values. The bootstrap approximation does not need to estimate the conditional sparsity. We found a promising performance of the new test in comparison with some natural competitors.

Acknowledgements

This study was supported by Project MTM2013-41383P from the Spanish Ministry of Economy and Competitiveness, as well as the European Regional Development Fund (ERDF). Support from the IAP network StUDyS from the Belgian Science Policy is also acknowledged. Work of M. Conde-Amboage was supported by a grant from Fundación Barrié and FPU grant AP2012-5047 from the Spanish Ministry of Education. V. Patilea acknowledges financial support from the research program *New Challenges for New Data* of LCL and Genes.

Bibliography

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. Annals of Statistics, 2, 760–777.

Conde-Amboage, M., Sánchez-Sellero, C. and González-Manteiga, W. (2015). A lack-of-fit test of quantile regression models with multiple covariates. Computational Statistics & Data Analysis, 88, 128–138.

Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression. Biometrika, 98, 995–999.

He, X. and Zhu, L.-X. (2003). A lack-of-fit test for quantile regression. Journal of the American Statistical Association, 98, 1013–1022.

Horowitz, J.L. and Spokoiny, V.G. (2002). An adaptive, rate-optimal test of linearity for median regression models. Journal of the American Statistical Association, 97, 822–835.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica, 46, 33–50.

Kong, E. and Xia, Y. (2012). A single-index quantile regression model and its estimation. Econometric Theory, 28, 730–768.

Maistre, S., Lavergne, P and Patilea, V. (2014). Powerful nonparametric checks for quantile regression. arXiv:1404.0216 [math.ST].

Patilea, V., Sánchez-Sellero, C. and Saumard, M. (2015). Testing the predictor effect on a functional response. Journal of the American Statistical Association, *forthcoming*.

Redden, D.T., Fernandez, J.R. and Allison, D.B. (2004). A simple significance test for quantile regression. Statistics in Medicine, 23, 2587–2597.

Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. Econometric Theory, 14, 123–138.