# An asymptotic lack-of-fit test for multiple quantile regression

*Mercedes Conde-Amboage*[1], *César Sánchez-Sellero*[2]

[1]mercedes.amboage@usc.es, Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

[2]cesar.sanchez@usc.es, Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

Let us suppose that a response random variable $Y$ may depend on several explanatory variables, denoted by a vector $X$. Then we can write $Y = q_\tau(X) + \varepsilon$, where $q_\tau$ is a regression function reflecting the effect of $X$ on $Y$ and $\varepsilon$ is an error term. In quantile regression, it is assumed that $P(\varepsilon < 0|X) = \tau$ with $\tau \in (0,1)$, and then $q_\tau(X)$ is the conditional $\tau$-quantile of $Y$ given $X$.

In this work we are going to propose a new method to test whether the regression function belongs to a parametric model, that is, to test a null hypothesis $H_0 : q_\tau \in \{q_\tau(\cdot,\theta) : \theta \in \Theta \subset R^q\}$ where $\theta$ is an unknown parameter that can be estimated from a sample of $(X,Y)$, say $(X_1,Y_1),\ldots,(X_n,Y_n)$.

The new test is based on the idea that under the null hypothesis the indicators $Z_i = I(Y_i - q_\tau(X_i,\hat{\theta}) \leq 0)$, $i \in \{1,\ldots,n\}$ should not depend on $X$, where $\hat{\theta}$ is an estimator of $\theta$. Adjusting a logistic regression model of these indicators on the gradient vector $\frac{\partial q_\tau(X_i,\hat{\theta})}{\partial \theta}$, and denoting the resulting fitted indicators by $\hat{Z}_i = \hat{\beta}' \frac{\partial q_\tau(X_i,\hat{\theta})}{\partial \theta}$, the residuals from this logistic fit satisfy $\sum_{i=1}^n \left(Z_i - \hat{Z}_i\right) \frac{\partial q_\tau(X_i,\hat{\theta})}{\partial \theta} = 0$.

The next step will be to consider some basis functions, $u_1(\cdot), u_2(\cdot), \ldots$, like orthonormalized polynomials or cosine functions, that span the space of all functions of $X$. Then, taking $r$ functions $u_1,\ldots,u_r$, a score statistic can be constructed as $U_r = \sum_{i=1}^n \left(Z_i - \hat{Z}_i\right) (u_1(X_i),\ldots,u_r(X_i))'$ whose squared norm, normalized by $\tau(1-\tau)$, gives $S_r = \frac{1}{\tau(1-\tau)} U_r' U_r$. Finally, our test statistic will be $T = \max_{r\in\{1,\ldots,R\}} (S_r - 2r)$, where $R$ is a sufficiently large bound for the number of basic functions. The null hypothesis will be rejected if $T$ takes on a large value.

We have obtained the asymptotic distribution of $T$ under the null hypothesis, which does not depend on unknown parameters, so no bootstrap or resampling techniques are required. Besides the computational efficiency of this asymptotic test, consistency of the test versus alternatives was proven and some simulations were carried out to compare its power with other competing tests. Note also that the proposed test is naturally adapted to continuous and categorical explanatory variables.

The new test is applied to check quantile regression models proposed in the literature to describe the effect of a number of explanatory variables on the infant birth weight. Observe that quantile regression models are more useful than mean regression models to describe the effect on low birth weights.

**Keywords:** quantile regression, lack-of-fit test.