Logistic quantile regression for bounded outcomes using a family of heavy-tailed distributions

Christian E. Galarza[a], Panpan Zhang[b] and Víctor H. Lachos[c*]

[a]*Departamento de Matemáticas, Escuela Superior Politécnica del Litoral, ESPOL, Ecuador*

[b]*Department of Biostatistics, Epidemiology and Informatics. University of Pennsylvania, USA*

[c]*Department of Statistics, University of Connecticut, USA*

Abstract

Mean regression model could be inadequate if the probability distribution of the observed responses is not symmetric. Under such situation, the quantile regression turns to be a more robust alternative for accommodating outliers and misspecification of the error distribution, since it characterizes the entire conditional distribution of the outcome variable. This paper proposes a robust logistic quantile regression model by using a logit link function along the EM-based algorithm for maximum likelihood estimation of the pth quantile regression parameters in Galarza et al. (2017). The aforementioned quantile regression (QR) model is built on a generalized class of skewed distributions which consists of skewed versions of normal, Student's t, Laplace, contaminated normal, slash, among other heavy-tailed distributions. We evaluate the performance of our proposal to accommodate bounded responses by investigating a synthetic dataset where we consider a full model including categorical and continuous covariates as well as several of its sub-models. For the full model, we compare our proposal with a non-parametric alternative from the so-called `quantreg R` package. The algorithm is implemented in the R package `lqr`, providing full estimation and inference for the parameters, automatic selection of best model, as well as simulation of envelope plots which are useful for assessing the goodness-of-fit.

Keywords: Bounded outcomes; Quantile regression model; EM algorithm; Scale mixtures of Normal distributions.

1 Introduction

Logistic regression (LR) models are commonly used for analyzing data with categorical responses. Besides, LR sometimes refers to a regression model which exploits a logistic link function to connect the systematic components of the model with appropriate transformations of the response variable. LR is popular due to its simple interpretations of the regression coefficients which quantify the change of the log-odds of one class over another with respect to each unit change in the associated covariate. On the other hand, logistic link function is often used for the transformation of an arbitrary number on the real line to a fixed number in the unit interval $[0, 1]$.

*Correspondence to: Department of Statistics, University of Connecticut, Storrs, CT 06269, U.S.A. E-mail address: hlachos@uconn.edu (V.H. Lachos)

Traditional model-fitting methods that do not account for the interval nature of the responses may lead to inaccurate results, such as confidence intervals outside of the parametric space or outbound predictions, etc. Some well-developed models, for instance, beta regression (Ferrari & Cribari-Neto 2004) and its extensions (e.g., augmented zero-one beta regression (Galvis et al. 2014) and mixture of beta distributions, Verkuilen & Smithson 2012) allow us to model the data with continuous responses whose values are in the unit interval. Beta regression is attractive for modeling this kind of data because unit-interval responses are U- or J-shaped in general (i.e., most observations appear as extremes in the interval). This particular feature can be captured by the parameters α and β (of beta distributions). For a general interval $[a,b] \in \Re$, the transformation to the unit interval can be simply done by standardization. Beta regression models thus can be applied to responses in any interval on the real line (see Ferrari & Cribari-Neto (2004) for example). This strategy has been commonly used in regression analysis owing to the lack of alternative models. Other models, similar to beta regression, are proposed upon the conditional mean of the response variable, which is usually a function of a set of covariates of interest (Gómez-Déniz et al. 2014, Paz et al. 2017). Nonetheless, all these models are subject to a series of constraints which must be verified.

One of the major drawbacks of the models mentioned above is that they are built upon the mean of the response variable. However, mean is not a robust measure for central tendency especially when the data is not symmetric, which is actually quite common for interval data. In addition, mean is not invariant to nonlinear transformations; that is, any nonlinear transformation of the mean of a variable is not identical to the mean of the variable under the same transformation, which results in that an appropriate model that fits the transformed data does not necessarily fit the original (untransformed) data. Hence, from a practical perspective, there is a need to seek a robust model that is not only resistant for data transformations, but also possesses appealing properties.

In this paper, we consider the QR model proposed by Koenker & G Bassett (1978). QR is more robust than many other competing mean-based models since

1. the model does not require any assumption for the error distribution;

2. the model is less sensitive to outliers.

As the QR model characterizes the conditional distribution of the response variable, it can be used to study the effect of covariates under different quantiles. The model is motivated by the natural idea of replacing mean by median as a central tendency measure when the response data is severely asymmetric. Another advantage of considering quantiles in the model is that quantiles preserve order, i.e., they are invariant to (monotonic) transformations. That means significance in transformed data always suggesting significance in original data. The QR model has become an attractive tool for dealing with bounded responses; see (Bottai et al. 2010) for instance. More recently, a fully Bayesian approach was discussed by Bayes et al. (2017) to model conditional quantiles of multivariate bounded responses, where a generalized double-bounded distribution proposed by Kumaraswamy (1980) was adopted as the link function.

It is mentioned in Bottai et al. (2010) that the QR model under the logistic transformation of the interval responses is similar to the proportional odds model for interval data (McCullagh 1980, 1984). In this paper, we emulate the work in Bottai et al. (2010) by considering a logistic transformation along a parametric QR model that considers a family of zero-quantile skewed distributions (SKD) proposed by Wichitaksorn et al. (2014). This class of distributions was also studied by Galarza et al. (2017), who introduced an EM-based quantile regression model considering the error to follow a member of this last class. It is worth mentioning that the latter generalizes

QR models based on the well-known asymmetric Laplace distribution (ALD) due to the ALD belongs to the SKD family. Furthermore, this parametric method showed a good recuperation of paramters in terms of precision and consistence of its ML estimates through extensive simulation studies. Due to the invariant nature of quantiles, we have that this good performance is passed to the interval response framework by simply considering any type of link function.

Next, we propose a logistic quantile regression model based on the parametric QR model before and a logit link function (due to its flexibility and its ease of interpretation).

The developed methodologies are coded in the `R lqr` package (Galarza et al. 2015), and ready to use. The built-in functions in this package not only allow for automatic fit of all the distributions presented in this paper, but also provide graphic tools for assessing the goodness-of-fit and likelihood-based criteria for model selection.

The rest of the paper is organized as follows. In Section 2, we give a brief review about QR, we introduce the novel class of skewed and heavy-tailed distributions generalized from a skewed normal as well as a EM algorithm for QR parameter estimation based on this class of distributions. In Section 3, we introduce a logistic QR model incorporating a logit link with the QR estimation model proposed in Section 2. We apply the proposed model to a synthetic data given in Galarza et al. (2015) under different covariate settings. We finally address some concluding remarks, and briefly discuss future work in Section 5. Some coding and figure outputs from our package are presented in the Appendix section.

2 Quantile Regression

In this section, we give a brief review about QR as preliminaries, and then we introduced the generalized class of skewed and heavy-tailed distributions that will be used throughout the remainder of the manuscript.

2.1 Introduction to quantile regression

Let $y_i,\ i=1,\ldots,n$, be observed responses, $\mathbf{x}_i$ be a dimension k-dimensional covariate vector for the ith subject, and $Q_p(y_i|\mathbf{x}_i)$ be the pth $(0<p<1)$ conditional quantile of y_i given $\mathbf{x}_i$. Assume that there is a linear relationship between the conditional quantile and the covariates $\mathbf{x}_i$, i.e., $Q_p(y_i|\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_p$, where $\boldsymbol{\beta}_p$ is an unknown k-dimensional parameter vector of primary interest. The quantile regression model is then given by

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_p + \varepsilon_i,\ i=1,\ldots,n, \tag{1}$$

where ε_i is the error term with density function $f_p(\cdot)$ such that its pth quantile is equal to zero, i.e., $\int_{-\infty}^{0} f_p(\varepsilon_i)d\varepsilon_i = p$. The pth QR estimate, denoted $\widehat{\boldsymbol{\beta}}_p$ can be obtained by minimizing the absolute sum of deviations (Koenker 2005):

$$\widehat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{\beta}_p\in\mathscr{R}^k}{\arg\min} \sum_{i=1}^{n} \rho_p\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_p\right), \tag{2}$$

where $\rho_p(\cdot)$ is the well-known *check function* given by $\rho_p(u) = u(p - \mathbb{I}\{u<0\})$. Specifically for $p=1/2$ (i.e., median regression), the check function can be simplified to $\rho_{1/2}(u) = |u|/2$; see Figure 1 for a graphic illustration.

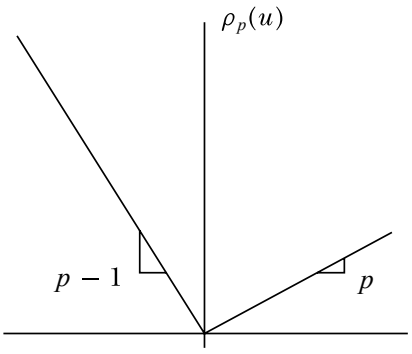

Figure 1: Check function for a given percentile p.

In general, the solution to Equation (2) does not have a closed form. Researchers usually utilize numerical methods to find the estimate for $\boldsymbol{\beta}_p$. That is, solving Equation (2) is equivalent to solving an L_1 optimization problem. One of the most well-known approaches is the *Lasso Penalized Quantile Regression (LPQR)*, a nonparametric method proposed by Tibshirani (1996); the main idea is to specify a penalty parameter that determines the shrinkage in the estimation process. An alternative algorithm is developed by Barrodale & Roberts (1977) to solve the linear programming problems proposed by Koenker & d'Orey (1987); the core of the algorithm is to compute the standard errors by using the *rank inversion* method proposed in Koenker (2005). Most of the parametric methods, no matter by frequentists (Benites et al. 2013, Zhou et al. n.d., Tian et al. 2014) or by Bayesians (Kottas & Gelfand 2001, Yu & Moyeed 2001, Kottas & Krnjajić 2009) are limited as they are applicable to only a few number of distributions. To the best of our knowledge, the *asymmetric Laplace distribution* (ALD) is the only well-defined distribution that possesses the zero-quantile property (i.e., $\int_{-\infty}^{0} f_p(\varepsilon_i)d\varepsilon_i = p$). However, the density function of ALD is not differentiable at zero, which may lead to numerical instability in computation. One remedy is to employ a convenient stochastic representation of ALD random variable, viz., a mixture of normal and exponential, hence allowing for the implementation of many other well-developed algorithms, such as the EM algorithm (Dempster et al. 1977) and its extensions.

2.2 A generalized class of skewed heavy-tailed densities

Owing to the deficiency of generating ADL random variables, the authors of Wichitaksorn et al. (2014) recently proposed a skewed normal (SKN) distribution possessing the zero-quantile property. Specifically, a random variable Y is said to follow an SKN distribution with location parameter μ, scale parameter $\sigma > 0$ and skewness parameter $p \in (0,1)$ if its probability density function (pdf) is in terms of

$$f(y|\mu,\sigma,p) = 2\left[p\,\phi\left(y\middle|\mu,\frac{\sigma^2}{4(1-p)^2}\right)\mathbb{I}\{y\le\mu\} + (1-p)\,\phi\left(y\middle|\mu,\frac{\sigma^2}{4p^2}\right)\mathbb{I}\{y>\mu\}\right], \qquad (3)$$

where $\phi(\cdot|\mu,\sigma^2)$ represents the pdf of the normal distribution with mean μ and variance σ^2 (i.e., $N(\mu,\sigma^2)$) and $\mathbb{I}\{A\}$ denotes the indicator function of set A. The density function in Equation (3) suggests that an SKD distribution are in essence a mixture of two truncated normal distributions. More recently, Galarza et al. (2017) adopted the check function to give a stochastic representation of SKN distribution, in terms of a class of scale mixtures of normal distributions (Andrews

& Mallows 1974), suggesting a family of heavy-tailed distributions with the zero-quantile property. We call this family of distributions *SKD family*. SKD family includes normal, Student's t, Laplace, slash, Pearson type VII and contaminated normal, etc.; distribution type depends on mixture random variable.

Formally, a random variable Y is said to be a member of SKD family with location parameter μ, scale parameter $\sigma > 0$ and skewness parameter $p \in (0,1)$, if its pdf can be represented by

$$f(y|\mu,\sigma,p,\boldsymbol{\nu}) = \int_0^\infty \frac{4p(1-p)}{\sqrt{2\pi\kappa(u)\sigma^2}} \exp\left\{-2\rho_p^2\left(\frac{y-\mu}{\kappa^{1/2}(u)\sigma}\right)\right\} dH(u|\boldsymbol{\nu}), \tag{4}$$

where $\kappa(\cdot)$ is a weight function and $h(u|\boldsymbol{\nu})$ is the pdf of mixture random variables, and $\boldsymbol{\nu}$ is a convenient vector of auxiliary parameters. Weight function and mixture random variable jointly determine the resultant distribution of Y. By convention, we shall write $Y \sim \text{SKD}(\mu,\sigma,p,\boldsymbol{\nu})$ if Y belongs to SKD family. The normal case is recovered when U is a degenerate random variable such that $P(U=1)=1$.

Several examples with specific choices of $\kappa(u)$ and $h(u|\boldsymbol{\nu})$ for the SKD family are shown in Table 1, and their associated density functions (taking $\mu=0$ and $\sigma=1$) are given in Figure 2. For each distribution in Figure 2, we choose three different values of p to show the differences in the shape and the skewness of density functions. We would like to point out that p, in fact, is the mass of the left tail of the pdf (owing to the zero-quantile property), and hence the density functions become symmetric when $p=1/2$.

Distribution	$\kappa(u)$	$h(u\|\boldsymbol{\nu})$	$f(y\|\mu,\sigma,\boldsymbol{\nu})$
Skewed Normal	u	$\mathbb{I}\{U=1\}$	$\frac{4p(1-p)}{\sqrt{2\pi\sigma^2}}\exp\left\{-2\rho_p^2\left(\frac{x-\mu}{\sigma}\right)\right\}$
Skewed Student's t	u^{-1}	$\text{G}(\frac{\nu}{2},\frac{\nu}{2})$	$\frac{4p(1-p)\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}}\left\{\frac{4}{\nu}\rho_p^2\left(\frac{y-\mu}{\sigma}\right)+1\right\}^{-\frac{\nu+1}{2}}$
Skewed Laplace	u	$\text{Exp}(2)$	$\frac{2p(1-p)}{\sigma}\exp\left\{-2\rho_p\left(\frac{y-\mu}{\sigma}\right)\right\}$
Skewed slash	u^{-1}	$\text{Beta}(\nu,1)$	$\nu\int_0^1 u^{\nu-1}\phi_{skd}(y\|\mu,u^{-1/2}\sigma,p)du$
Skewed cont. normal	u^{-1}	$\nu\mathbb{I}\{u=\gamma\}+(1-\nu)\mathbb{I}\{u=1\}$ $0\le\nu,\gamma\le 1,$	$\nu\phi_{skd}(y\|\mu,\gamma^{-1/2}\sigma,p)+(1-\nu)\phi_{skd}(y\|\mu,\sigma,p)$

Table 1: $\kappa(\cdot)$, $h(u|\boldsymbol{\nu})$ and pdfs for some members of the SKD family. $\text{Exp}(\beta)$ denotes the exponential distribution with mean β, and $\phi_{skd}(y|\mu,\sigma,p)$ denotes the density function in Equation (3).

In this subsection, we present the estimation method for the QR model in Galarza et al. (2017), which considers an error term belong to the class of skewed distributions described before.

2.3 Parameter estimation via the EM algorithm

Using the hierarchical representation of a SKD variate, the QR model defined in (1) can be expressed as

$$\begin{aligned} Y_i|U_i=u_i &\sim \text{SKN}(\mathbf{x}_i^\top\boldsymbol{\beta}_p, \sqrt{\kappa(u_i)}\sigma, p), \\ U_i &\sim h(u_i|\boldsymbol{\nu}), \end{aligned}$$

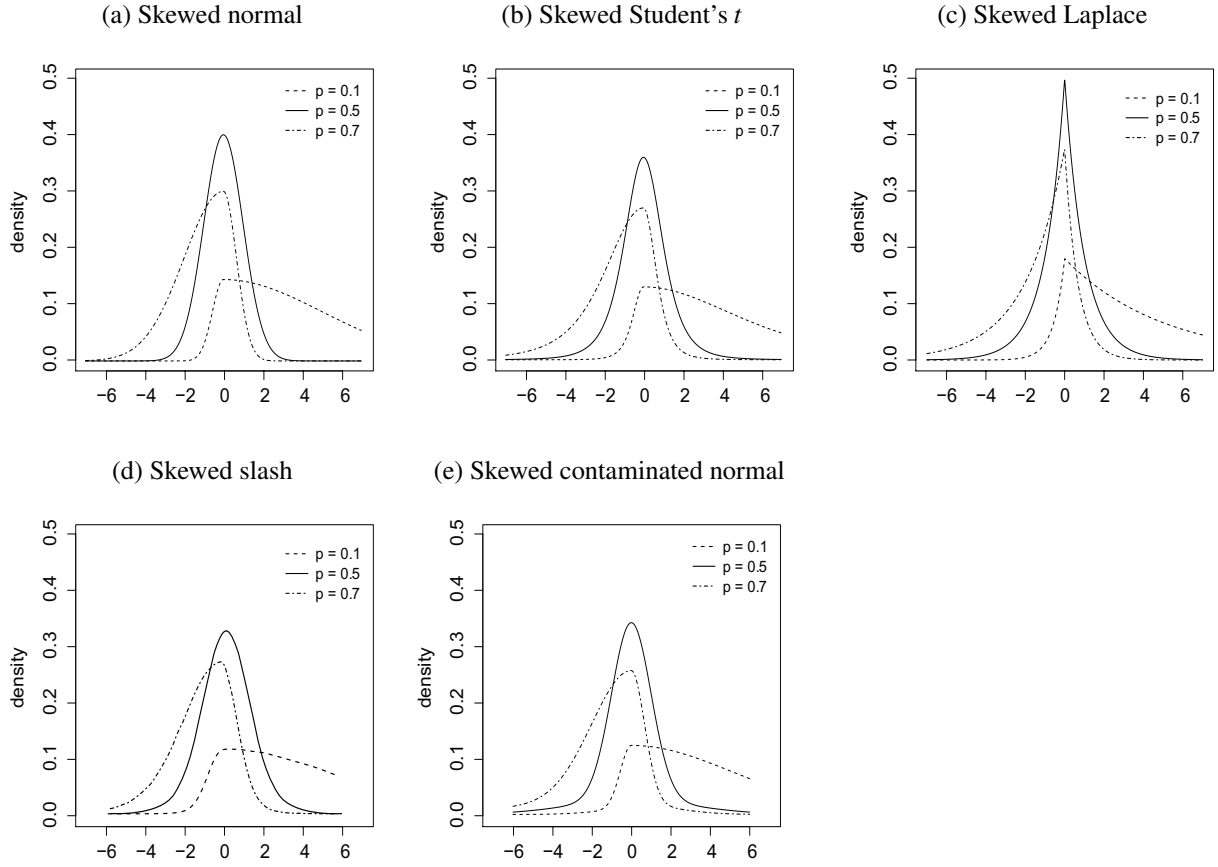

Figure 2: Density functions for the standard skewed normal, skewed Student's t ($\nu = 4$), skewed Laplace, skewed slash ($\nu = 2$) and skewed contaminated normal ($\nu = 4$ and $\gamma = 0.1$) distributions under different values of skewness parameter.

where $h(u|\boldsymbol{\nu})$ represents the mixture density. Let $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{u} = (u_1, \ldots, u_n)$ be the observed and missing (latent) data, respectively. Then, the complete data log-likelihood function $\ell_c(\boldsymbol{\theta}|y_i, u_i)$ can be rewritten as

$$\ell_c(\boldsymbol{\theta}|y_i, u_i) = \sum_{i=1}^{n} \log \phi \left(y_i \left| \mathbf{x}_i^\top \boldsymbol{\beta}_p, \frac{\kappa(u_i)\sigma^2}{4\xi_i^2} \right. \right) + \sum_{i=1}^{n} \log h(u_i|\boldsymbol{\nu}),$$

for $i = 1, \ldots, n$, and where $\xi_i = (1-p)\,\mathbb{I}\{y_i \leq \mathbf{x}_i^\top \boldsymbol{\beta}_p\} + p\,\mathbb{I}\{y_i > \mathbf{x}_i^\top \boldsymbol{\beta}_p\}$.

The E step of the EM algorithm requires evaluation of the so-called Q-function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \mathrm{E}[\ell_c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u})|\mathbf{y}, \boldsymbol{\theta}^{(k)}]$; however, it is required to compute $\widehat{\kappa^{-1}(u_i)} = \mathrm{E}[\kappa^{-1}(U_i)|y_i, \boldsymbol{\theta}^{(k)}]$, that will depend of the weight function $\kappa(\cdot)$. Details of particular cases for this expectation are shown in the Appendix section. In what follows the superscript (k) will indicate the estimate of the related parameter at the stage k of the algorithm.

The proposed EM algorithm can be summarized in the following steps:

1. ***E-step***: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, compute $\widehat{\kappa^{-1}(u_i)}$.

2. ***M-step***: Update $\boldsymbol{\theta}^{(k)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ over $\boldsymbol{\theta}$, which leads to the following expres-

sions

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_p^{(k+1)} &= (\mathbf{X}^\top \boldsymbol{\Omega}^{(k)} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{(k)} \mathbf{y}, \\ \widehat{\sigma^2}^{(k+1)} &= \frac{4}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_p^{(k+1)})^\top \boldsymbol{\Omega}^{(k)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_p^{(k+1)}),\end{aligned}$$

where $\boldsymbol{\Omega}$ is a $n \times n$ diagonal matrix, with elements $\xi_i^2 \widehat{\kappa^{-1}(u_i)}$, $i = 1, \ldots, n$, $\mathbf{X}$ is the design matrix and $\mathbf{y}$ is the vector of observations. After the *M-step*, we will update the parameter $\boldsymbol{\nu}$ by maximizing the marginal log-likelihood function of $\mathbf{y}$, obtaining

$$\widehat{\boldsymbol{\nu}}^{(k+1)} = \arg\max_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log f(y_i | \widehat{\boldsymbol{\beta}}_p^{(k+1)}, \widehat{\sigma}^{(k+1)}, \boldsymbol{\nu}).$$

In practice, the EM algorithm iterates until some distance involving two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta})$, like $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})||$ or $||\ell(\boldsymbol{\theta}^{(k+1)})/\ell(\boldsymbol{\theta}^{(k)}) - 1||$, is small enough. We have use ordinary least squares estimators (OLSE) as an initial estimate of $\boldsymbol{\beta}$, reaching convergence in a few seconds.

The QR model based on the EM algorithm above showed to have a good performance in terms of precision and consistence of its ML estimates, this being assessed through a full simulation study. Then, in order to inherit these properties, it is enough to use any transformation from $\mathscr{R} \to [a,b] \subset \mathscr{R}$, say, a link function, in order to fit bounded responses. Next, we propose a logistic quantile regression model based on the parametric QR model before and a logit link function (due to its flexibility and its ease of interpretation).

3 Logistic quantile regression

Let y be an interval response taking values in $[a,b] \subset \mathscr{R}$ such that $a < b$. The authors of Bottai et al. (2010) introduced a standardized logistic transformation for an arbitrary interval response as follows:

$$h(y) = \log\left(\frac{y-a}{b-y}\right),$$

where a and b are the lower and upper bound limits, respectively. Note that $y-a$ and $b-y$ are both positive quantities, and they respectively respresent the distances from y to its lower and upper limits. Thus, this ratio is always positive. To ensure that the logistic transformation is well defined for all values of y, it is recommended to add a small perturbation to the bounds, say, $a^* = a - \varepsilon$ and $b^* = b + \varepsilon$, where ε is a reasonably small quantity.

For the ith subject, its pth conditional quantile can be obtained by using the inverse standardized logistic transformation; namely,

$$Q_p(y_i|\mathbf{x}_i) = \frac{b\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_p) + a}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_p)}. \tag{5}$$

Note that Equation (5) coincides with the traditional expression of logistic mean regression for $a = 0$ and $b = 1$. Regardless of the quantiles to be predicted, this transformation recovers the estimation of the fixed effects for the original responses due to the invariance property of quantiles to monotonic transformations. According to the fundamental property of distribution

function, we know that, for any random variable Y and any nondecreasing function h, we have $P(Y \leq y) = P(h(Y) \leq h(y))$, and hence

$$Q_p(h(y)|\mathbf{x}_i) = h(Q_p(y_i|\mathbf{x}_i)).$$

This is a standard technique in a great deal of work on QR under different kinds of monotonic transformations (see for example, Powell 1986, Mu & He 2007). However, we would like to remind the readers that this property does not hold for mean-based regression, as, for instance, the average of logarithms does not equal the logarithm of the averages in general. Specifically, if the distributions of x and y, both measured on logarithmic scales, are normal, then the arithmetic mean regression of $\log(y)$ on $\log(x)$ is linear; that is, the geometric mean regression of y on x is in the form of $y = \beta_0 x^{\beta_1}$, where β_0 and β_1 are constants.

Although other link functions could be considered (e.g., `probit`, `log-log`, `complementary log-log`, etc.), we focus only on `logit` link function for the sake of interpretation of the parameters. We can write the inverse of Equation (5) as follows:

$$\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_p) = Q_p\left(\frac{y_i - a}{b - y_i}\middle|\mathbf{x}_i\right) = \frac{Q_p(y_i|\mathbf{x}_i) - a}{b - Q_p(y_i|\mathbf{x}_i)}, \tag{6}$$

which converges to $p/(1-p)$ as $n \to \infty$. Equation (6) is obtained by applying the invariant property of quantiles to Equation (5), best reflecting the flexibility of QR models for transformed variables. Besides, $\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_p)$ can be viewed as the odds ratio of an observation less than or equal to $Q_p(y_i|\mathbf{x}_i)$, suggesting that its interpretation is very much similar to binary regression model. As noted, the asymptotically probability that an observation falls in the interval $[0, Q_p(y_i|\mathbf{x}_i) - a]$ is equal to $P(Y \leq Q_p(y_i|\mathbf{x}_i)) = p$. Hence, Equation (6), on the other hand, is related to the proportional odds model for interval responses; see McCullagh & Nelder (1989). The utilization of the model in Equation (6) is to study responses in the following two intervals: below the pth quantile (primary interest in this paper) and above the pth quantile. In most cases, the model shows monotonicity with respect to intercepts (in view of proportional odds models); namely,

$$\beta_{01} \leq \beta_{02} \leq \ldots \leq \beta_{0k}, \tag{7}$$

where β_{0j} is the intercept of the QR model for the p_jth quantile, with $(p_1, p_2, \ldots, p_k)$ representing an ordered grid of probabilities. The verification of Equation (7) is straightforward, as, by definition, we have $Q_{p_1}(y) \leq Q_{p_2}(y) \leq \ldots \leq Q_{p_k}(y)$. We would like to mention that Equation (7) always hold for proportional odds models, since all the intercepts are estimated simultaneously. However, it may not be always true for QR models, since each quantile is estimated separately. For example, Equation (7) may be invalid when the sample size is small. In the last decade, several nonparametric methods were proposed to estimate the quantiles simultaneously. We refer the interested readers to Liu & Wu (2009, 2011) for multiple non-crossing QR models.

4 Applications

In this section, we present the application of the proposed model to a synthetic dataset "tumor cell resistance to death," which is available in the `R lqr` package. The experiment measures the resistance to death of two types of tumors, say A and B (`type`), under different doses of an experimental drug. The dose of the drug is a positive random quantity (`dose`) and the response variable is a score taking values from 0 to 4, where 0 represents the greatest resistance to death (`score`) . We

follow the research outline in Bottai et al. (2010). Our goal is to properly interpret the parameter estimates for three models under different covariate settings: a model with a categorical covariate, a model with continuous covariates only, and a full model with both categorical and continuous covariates.

4.1 A categorical covariate

The scientific question in which we are interested is what type of tumor influences more on the resistance to death and how it works. The boxplots in Figure 3 shows that the scores are different between two types of tumor cells. Both distributions are right-skewed, indicating the lack of normality (The nonparametric *Shapiro-Wilk* test shows the p-values less than 10^{-7} for both types). The difference in group means is 0.41 units, while the difference of group medians is 0.57 units. It appears that type B tumor cells are less resistant with higher scores. Since the distributions are skewed, median appears to be a more appropriate measure for central tendency. We consider the following model:

$$\log\left(\frac{\texttt{score}_i+\varepsilon}{4-\texttt{score}_i+\varepsilon}\right)=\beta_0+\beta_1\texttt{type}_i+\varepsilon_i,$$

where `type` is $(0=\text{Type A}, 1=\text{Type B})$ and the error following a distribution from the SKD family, i.e., $\varepsilon_i\sim \text{SKD}(\mu,\sigma,p,\boldsymbol{\nu})$. To guarantee that the fraction is well defined for $\texttt{score}=0$ or $\texttt{score}=4$, we consider a small quantity $\varepsilon=0.001$ to the bounds.

Figure 4 shows the output for the Student's t model, which was the best fit. From this output, we confirm that there is a significant difference between the two group medians. We also observe that the group difference is more significant for lower quantiles, but the gap is narrowing with the increase of scores. In addition, it seems that the two tumor types have approximately equal resistance to death for quantiles 0.75 and 0.95. We infer that 25% of tumors with the lowest resistance to the drug (i.e., highest score values) are from both types. It seems necessary to study the magnitude of the effect of tumor types on the resistance. Note that

$$\frac{Q_p(\text{odd}(\texttt{score}_A))}{Q_p(\text{odd}(\texttt{score}_B))}=\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)}=\exp(\beta_1).$$

Then, $\exp(\beta_1)=\exp(0.63)$ is the odds ratio of the median score between tumor types B and A. In the saturated model. This corresponds to the observed odds ratio

$$\frac{1.66/(4-1.66)}{1.09/(4-1.09)}=\exp(0.63)=1.89,$$

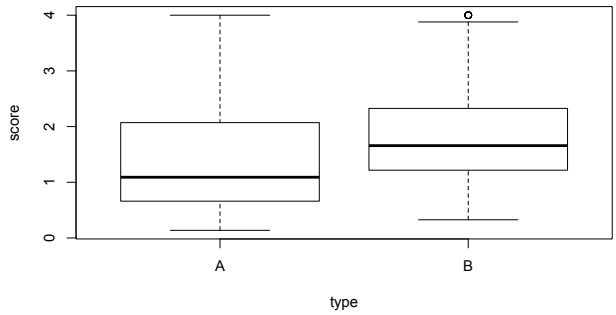

Figure 3: Side by side boxplots of the resistance to death score of two types of tumors.

where 1.66 and 1.09 are the observed median scores for tumor types B and A, respectively. We also apply the LQ regression for binary variables. By studying the effects signal, we find that the median of type A tumor (the reference group) is greater than the median for type B tumor only for the 0.95 percentile; however, the difference is not significant for 90% confidence. Figure 5 shows a study of the fixed effects of tumor type along with different quantiles. As expected, the intercept always increases as quantile value increases. As of the type effect, the odds ratio for the 0.05 quantile is approximately 3 and decays until it is no longer significant when the quantile value exceeds 0.70.

```
Quantiles
             0.05       0.25       0.50       0.75      0.95
Intercept -3.6378    -2.5137    -1.6380    -0.4871    1.8379
Type(B)    1.1017***  0.8417***  0.6325**   0.3697.   -0.1618
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Figure 4: Sample output from `R lqr` package using the Student's t model.

Besides, LQ regression can be used to compare multiple groups by adding dummy variables into the model. The best-fit model (automatically chosen by the `Log.best.lqr` function) for the median regression is the one with error term following the Student's t distribution with $\nu = 2$, indicating a significant departure from normal by heavy tails. Histograms of the residuals for all listed SKD models and associated output codes can be found in the Appendix.

4.2 A continuous covariate

Next, we model the quantiles of the resistance to death as a function of the amount of drug supplied. The scatterplot in Figure 6 implies that the higher the dose is, the higher the score is, and consequently the tumor dies more easily. Conditional boxplots show that the quantiles increase as the dose increases, but the increments vary from quantile to quantile. The relationship between dose and score is positive but not linear. Noticing that the data pattern presents an apparent inflection point, we propose and compare the following two cubic models:

$$\texttt{score}_i = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 + \beta_3 \texttt{dose}_i^3 + \varepsilon_i \tag{8}$$

and

$$\log\left(\frac{\texttt{score}_i + \varepsilon}{4 - \texttt{score}_i + \varepsilon}\right) = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 + \beta_3 \texttt{dose}_i^3 + \varepsilon_i, \tag{9}$$

where the model in Display (8) does not consider the interval nature of response. The error term (ε_i) in both models and the small quantity ε in Model (9) are set in the same manner as in Subsection 4.1. For both models, we fit a QR for quantiles $p = \{0.05, 0.50, 0.95\}$. Figure 7 shows the difference between adjusted quantiles for a traditional model and those for the proposed LQ regression model. We observe that the predicted scores from the traditional model are not bounded within the domain interval, $[0,4]$. We conclude that the predictions of the scores for some extreme values of quantiles may go beyond the bounds, when the models not accounting for the constraints (of the interval data) are used. Indeed, these models are not feasible to the study.

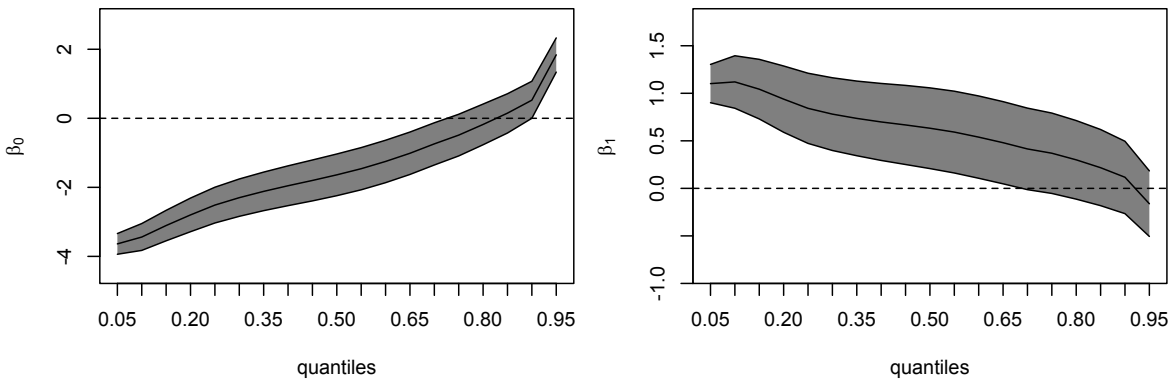

Figure 5: Point estimates and 95% confidence interval for the fixed effects for the set of quantiles $p = \{0.05, \ldots, 0.95\}$ under a Student's t model. Graphs are generated by the `Log.lqr` function.

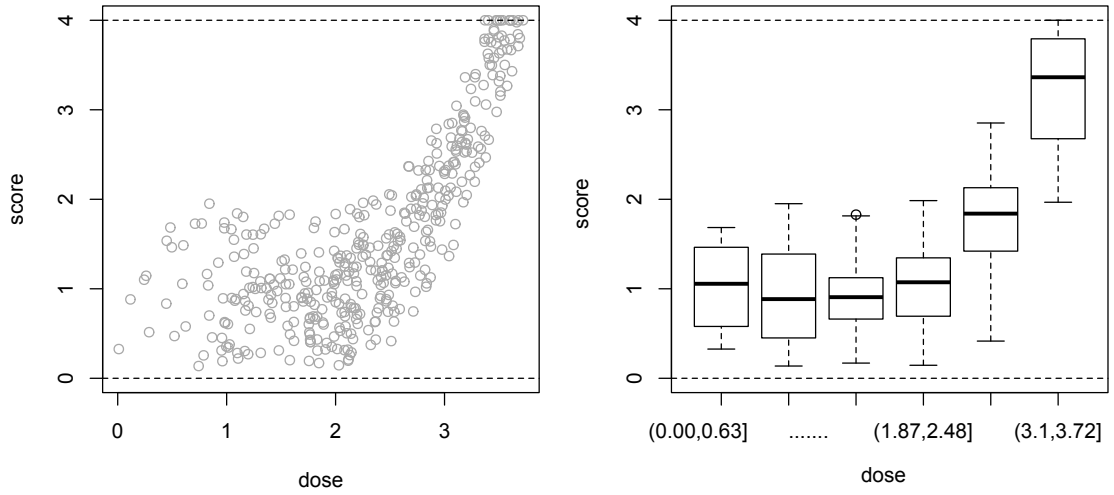

Figure 6: Descriptive summaries of the tumor cell resistance to death dataset.

For any quantitative covariate x, we have

$$\frac{\text{odd}(Q_p(Y|x+1))}{\text{odd}(Q_p(Y|x))} = \frac{\exp(\beta_0+\beta_1(x+1))}{\exp(\beta_0+\beta_1 x)} = \exp(\beta_1),$$

and conclude that for each unit increase of x, we expect a $\exp(\beta_1)\times 100\%$ increase (or decrease) on the odds ratio of observing a value less than $Q_p(Y)$. In addition, we observe that the cubic effect of dose is significant at 95% confidence for all quantiles. The best-fit model turns to be a heavy-tailed slash distribution with $\nu \leq 4$. A complete study including categorical and continuous variables will be conducted in the next subsection.

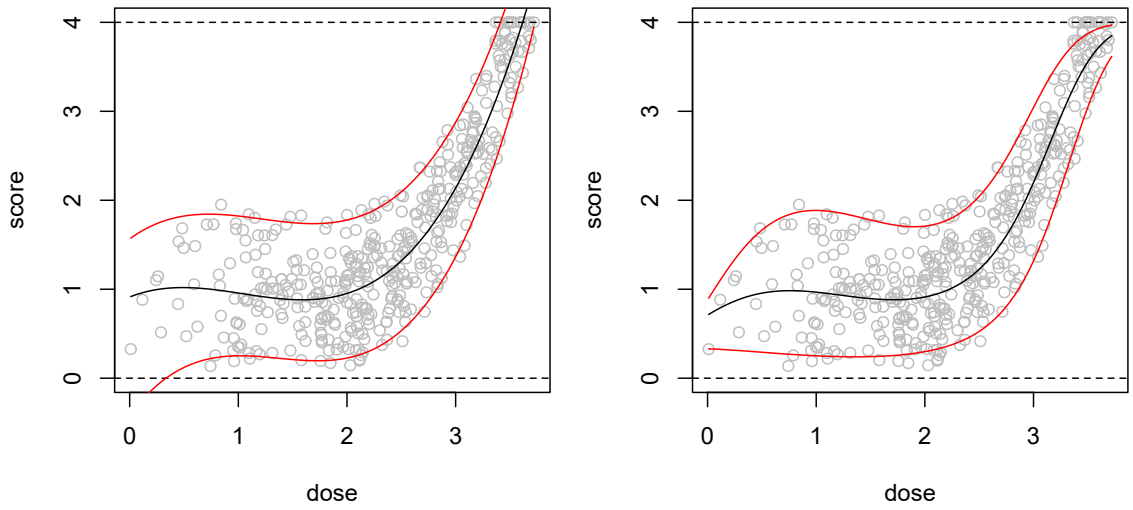

Figure 7: Fitted prediction curves using a traditional model and a logistic quantile regression for the scores of the resistance to death under a slash model. The quantiles are set at $p = \{0.05, 0.50, 0.95\}$. The median predicted curves are shown in black, bounded by red solid curves for the 5th and 95th quantiles, respectively.

4.3 Full model

Finally, we propose a full LQ model as follows:

$$\log\left(\frac{\texttt{score}_i + \varepsilon}{4 - \texttt{score}_i + \varepsilon}\right) = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 + \beta_3 \texttt{dose}_i^3 + \beta_4 \texttt{type}_i + \beta_5 (\texttt{type} * \texttt{dose})_i + \varepsilon_i,$$

in which we consider the cubic dose on the score, the effect of tumor type and the interaction of tumor type and dose simultaneously. We investigate the model fit for a dense set of quantiles $p = \{0.05, 0.10, \ldots, 0.95\}$. We fit the full model through the function `Log.best.lqr` from our proposed package `lqr`, along with the likelihood-based criteria, and observe that the slash distribution provides the best fit (see Table 2) between five tested models. For all quantiles, we get the proper degrees of freedom $\nu \leq 2$, suggesting that the chosen distribution is heavy-tailed, and that the proposed model is more appropriate to the study owing to its robustness. According to Table 2, the second best choice is the Student's t distribution.

	Normal	Student's t	Laplace	Slash	C. Normal
AIC	1835.3102	933.1516	1027.4132	915.4108	1165.5846
BIC	1863.6748	961.5162	1055.7778	943.7754	1193.9492
HQ	1846.5158	944.3572	1038.6189	926.6164	1176.7902
loglik	-910.6551	-459.5758	-506.7066	-450.7054	-575.7923

Table 2: Likelihood-based model selection criteria for the logistic quantile regression for the median score of the resistance to death under all SKD models.

For sake of comparison, we also fit a logistic non-parametric quantile regression by incorporating a logit link function along the `qr` function from the so called `quantreg R` package. In contrast with our proposal, this function uses a direct ℓ_1 optimization and by default it invokes a variant of the Barrodale and Roberts simplex algorithm described in Koenker & d'Orey (1987).

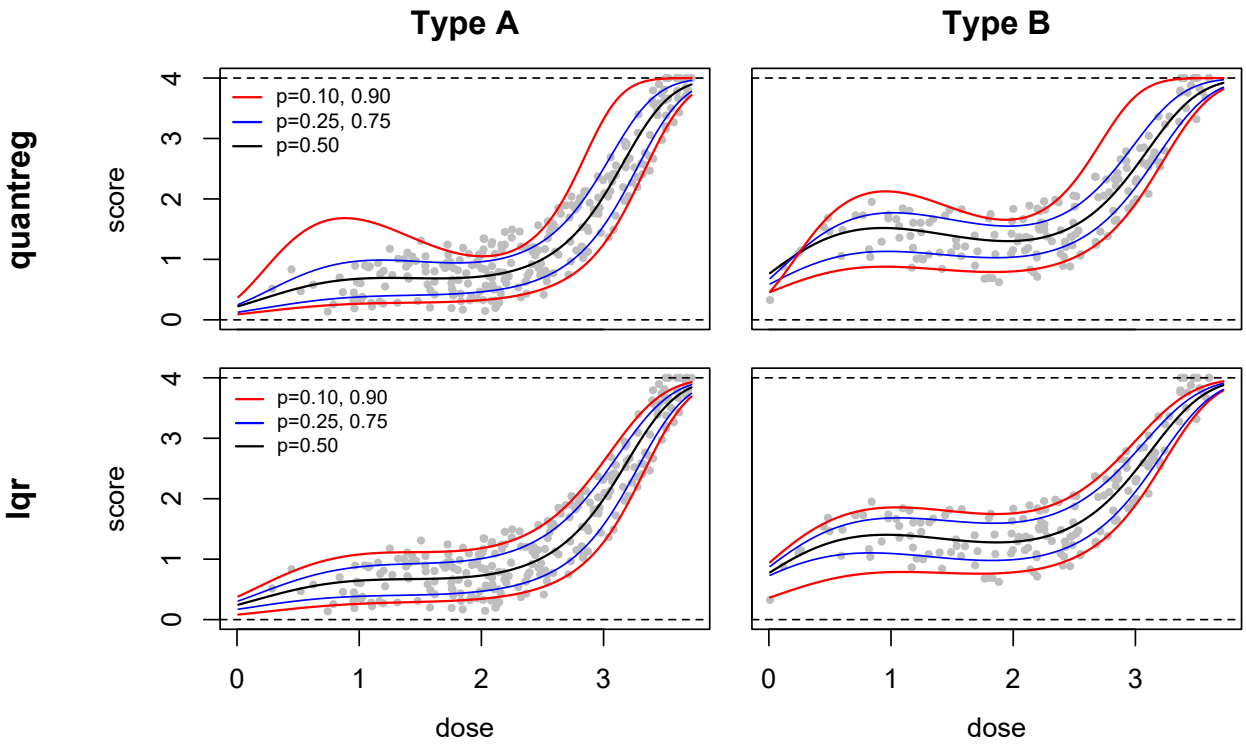

Figure 8: Fitted prediction curves using a logistic quantile regression of the score of the resistance to death for a grid of quantiles, different tumor types and under two models.

Figure 8 shows the fitted prediction curves for the full model for the quantiles $p = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ for the two tumor types (type A on the left and type B on the right), under both models (`quantreg` at top and our proposal `lqr` at bottom). To summarize, this example let us see how QR model characterizes the entire conditional distribution of the response variables. Note that our proposal model outperforms the non-parametric alternative, providing a better *enveloping* of the data. For instance, we see that the `quantreg` model has problems to deal with the high quantile 0.90, where the latter even crosses other quantile curves violating the monotonicity assumption. Although the QR model we use does not consider this restriction, it has proven to be robust enough to minimize this problem from occurring. Without a doubt, for the non-parametric model, the flexibility of the curves can be controlled with some smoothing parameter, however, specific adjustments would be far from the objective of comparing user-ready functions.

Finally, Figure 9 presents the point estimates and 95% confidence intervals (CIs) for the fixed effects $\boldsymbol{\beta}$ for the dense set of quantiles $p = \{0.05, \ldots, 0.95\}$ for both methods. Gray intervals represents the CIs provided by our `Log.lqr` function, while red intervals are the ones provided by the `rq` function from the `quantreg R` package. A quick glance let us to see that both methods offer significantly different results; however, a evidence that our method offers a better recovery of parameters is that the analyzed data set was generated using a zero-intercept cubic model with the cell type being significant (Galarza et al. 2015), and our proposal found the intercept to be 0 for all quantiles (zero line crosses the 95% CI across quantiles), being consistent with the nature of data. As expected, point estimates for intercept increase as quantiles increase as well. Furthermore, under our slash model, the effects of linear, quadratic and cubic dose are all significant in quantiles. Based on the fact that effect of dose is significant, increasing and positive, we conclude that the higher the dose is, the lower the resistance to death is for all quantiles, implying the experimental

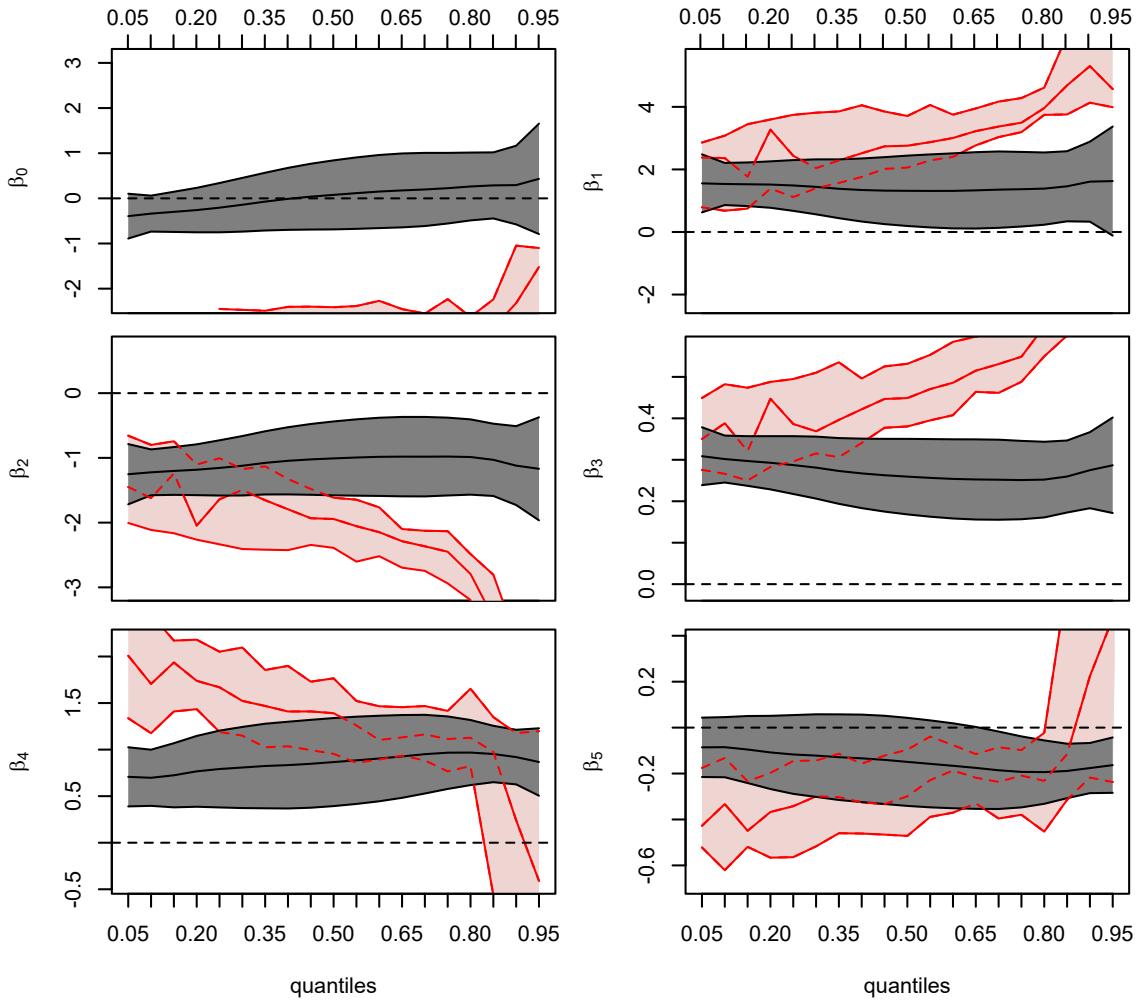

Figure 9: Point estimates and 95% confidence interval for the fixed effects $\boldsymbol{\beta}$ for the dense set of quantiles $p = \{0.05, \ldots, 0.95\}$. Gray and red intervals provided by `R Log.lqr` and `rq` functions, respectively.

drug does fulfill its predicted goal. In addition, the effect of the drug is inversely proportional to the resistance of the tumor, as expected. The effect of tumor type is significant for all quantiles, different from the results of the model considered in Subsection 4.1. The positivity of this effect indicates that tumor type A is more resistant to the experimental drug than type B for all quantiles, remaining almost constant along quantiles. The interaction between tumor type and dose seems to be significant. Since the 95% band is very close to zero for middle quantiles, we conclude that the interaction is more decisive for extreme quantiles where the CI shrinks.

5 Conclusions

In this paper, we propose a logistic quantile regression model to fit bounded responses. We combined the robustness of a parametric QR model that considers heavy-tailed errors, with the flexibility of a logit link function, providing a flexible robust model for estimating interval or

strictly positive responses, with ease of interpretation as showed in the Application section and being available via user-friendly functions from the `lqr R` package. The main function for LQR offers automatic selection of the best model, full inference, residual and envelope plots for assessing the goodness-of-fit, as well as CIs plots for a grid of quantiles. The package has been downloaded for almost 16K times since its release. We decide to make the codes of our method open to the public so as to encourage the interested researchers using the proposed model in their QR research.

In theory, the transformation methodology considered in this paper can be effortlessly extended to other existing QR models, such as, mixed effects models for longitudinal data and interval censored responses which are quite common in many areas like biology, medicine and pharmacology.

References

Andrews, D. F. & Mallows, C. L. (1974), 'Scale mixtures of normal distributions', *Journal of the Royal Statistical Society, Series B.* **36**, 99–102.

Barndorff-Nielsen, O. E. & Shephard, N. (2001), 'Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics', *Journal of the Royal Statistical Society, Series B* **63**, 167–241.

Barrodale, I. & Roberts, F. (1977), 'Algorithms for restricted least absolute value estimation', *Communications in Statistics-Simulation and Computation* **6**, 353–363.

Bayes, C. L., Bazan, J. L. & De Castro, M. (2017), 'A quantile parametric mixed regression model for bounded response variables', *Statistics and its Interface* **10**, 483–493.

Benites, L., Lachos, V. H. & Vilca, F. (2013), Likelihood based inference for quantile regression using the asymmetric Laplace distribution, Technical Report 15, Universidade Estadual de Campinas.

Bottai, M., Cai, B. & McKeown, R. E. (2010), 'Logistic quantile regression for bounded outcomes', *Statistics in Medicine* **29**(2), 309–317.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Ferrari, S. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**(7), 799–815.

Galarza, C. E., Benites, L. & Lachos, V. H. (2015), *lqr: Robust Linear Quantile Regression*. R package version 1.5.

Galarza, C., Lachos, V., Barbosa Cabral, C. & Castro Cepero, L. (2017), 'Robust quantile regression using a generalized class of skewed distributions', *Stat* **6**(1).

Galvis, D. M., Bandyopadhyay, D. & Lachos, V. H. (2014), 'Augmented mixed beta regression models for periodontal proportion data', *Statistics in Medicine* **33**(21), 3759–3771.

Gómez-Déniz, E., Sordo, M. A. & Calderín-Ojeda, E. (2014), 'The log–lindley distribution as an alternative to the beta regression model with applications in insurance', *Insurance: mathematics and Economics* **54**, 49–57.

Koenker, R. (2005), *Quantile Regression*, Vol. 38, Cambridge University Press.

Koenker, R. & G Bassett, J. (1978), 'Regression quantiles', *Econometrica: Journal of the Econometric Society* **46**, 33–50.

Koenker, R. W. & d'Orey, V. (1987), 'Algorithm as 229: Computing regression quantiles', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **36**(3), 383–393.

Kottas, A. & Gelfand, A. E. (2001), 'Bayesian semiparametric median regression modeling', *Journal of the American Statistical Association* **96**, 1458–1468.

Kottas, A. & Krnjajić, M. (2009), 'Bayesian semiparametric modelling in quantile regression', *Scandinavian Journal of Statistics* **36**, 297–319.

Kumaraswamy, P. (1980), 'A generalized probability density function for double-bounded random processes', *Journal of Hydrology* **46**(1-2), 79–88.

Liu, Y. & Wu, Y. (2009), 'Stepwise multiple quantile regression estimation using non-crossing constraints', *Statistics and its Interface* **2**(3), 299–310.

Liu, Y. & Wu, Y. (2011), 'Simultaneous multiple non-crossing quantile regression estimation using kernel constraints', *Journal of Nonparametric Statistics* **23**(2), 415–437.

McCullagh, P. (1980), 'Regression models for ordinal data', *Journal of the Royal Statistical Society. Series B (Methodological)* **42**(2), 109–142.

McCullagh, P. (1984), 'Generalized linear models', *European Journal of Operational Research* **16**(3), 285–292.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman & Hall/CRC.

Mu, Y. & He, X. (2007), 'Power transformation toward a linear regression quantile', *Journal of the American Statistical Association* **102**(477), 269–279.

Paz, R. F. d. et al. (2017), 'Alternative regression models to beta distribution under bayesian approach'.

Powell, J. L. (1986), 'Censored regression quantiles', *Journal of Econometrics* **32**(1), 143–155.

Tian, Y., Tian, M. & Zhu, Q. (2014), 'Linear Quantile Regression Based on EM Algorithm', *Communications in Statistics - Theory and Methods* **43**(16), 3464–3484.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* pp. 267–288.

Verkuilen, J. & Smithson, M. (2012), 'Mixed and mixture regression models for continuous bounded responses using the beta distribution', *Journal of Educational and Behavioral Statistics* **37**(1), 82–113.

Wichitaksorn, N., Choy, S. & Gerlach, R. (2014), 'A generalized class of skew distributions and associated robust quantile regression models', *Canadian Journal of Statistics* **42**(4), 579–596.

Yu, K. & Moyeed, R. (2001), 'Bayesian quantile regression', *Statistics & Probability Letters* **54**, 437–447.

Zhou, Y.-h., Ni, Z.-x. & Li, Y. (n.d.), 'Quantile Regression via the EM Algorithm', *Communications in Statistics - Simulation and Computation* (10), 2162–2172.

Appendix

Details of expectations in EM algorithm

The conditional distribution of the latent variable given the observed data $f(u_i|y_i,\boldsymbol{\theta}^{(k)})$ will depend on the functional form of $h(u_i|\boldsymbol{\nu})$. Table 3 shows the conditional *pdf* of U given Y for specific choices of $h(u_i|\boldsymbol{\nu})$.

Table 3: Conditional distribution of U given Y for specific SKD distributions.

Distribution	Distribution of U	Conditional distribution of $U\|Y$	$\widehat{\kappa^{-1}(u_i)}$
skewed Student-t	$\mathrm{G}(\frac{\nu}{2},\frac{\nu}{2})$	$\mathrm{G}\left(\frac{\nu+1}{2},\frac{\nu+4\xi_i^2z_i^2}{2}\right)$	$\frac{\nu+1}{\nu+4\xi_i^2z_i^2}$
skewed Laplace	$\mathrm{Exp}(2)$	$\mathrm{GIG}\left(\frac{1}{2},2\xi_i^2z_i^2,\frac{1}{2}\right)$	$\frac{1}{2\xi_i\lvert z_i\rvert}$
skewed slash	$\mathrm{Beta}(\nu,1)$	$\mathrm{TG}\left(\nu+\frac{1}{2},2\xi_i^2z_i^2,1\right)$	$\left[\frac{\nu+\frac{1}{2}}{2\xi_i^2z_i^2}\right]\frac{\mathscr{F}(1\|\nu+\frac{3}{2},2\xi_i^2z_i^2)}{\mathscr{F}(1\|\nu+\frac{1}{2},2\xi_i^2z_i^2)}$
skewed cont. normal	$\nu\mathbb{I}\{u=\gamma\}+(1-\nu)\mathbb{I}\{u=1\}$ $0\leq\nu,\gamma\leq1$	$\frac{a\mathbb{I}\{u=\gamma\}+b\mathbb{I}\{u=1\}}{a+b}$	$\frac{a\gamma+b}{a+b}$

In Table 3, $z_i=(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_p)/\sigma$ and $\mathscr{F}(x|\alpha,\lambda)$ represents the *cdf* of a Gamma (α,λ) distribution. Moreover, expressions for a and b are given by $a=\nu\phi\left(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}_p,\frac{\gamma^{-1}\sigma^2}{4\xi_i^2}\right)$ and $b=(1-\nu)\phi\left(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}_p,\frac{\sigma^2}{4\xi_i^2}\right)$. The notation $\mathrm{TG}(\alpha,\lambda,t)$ represents a random variable with $\mathrm{Gamma}(\alpha,\lambda)$ distribution truncated to the right at the value t. Finally, $GIG(\nu,a,b)$ denotes the Generalized Inverse Gaussian (GIG) distribution (see Barndorff-Nielsen & Shephard (2001) for more details).

Figures

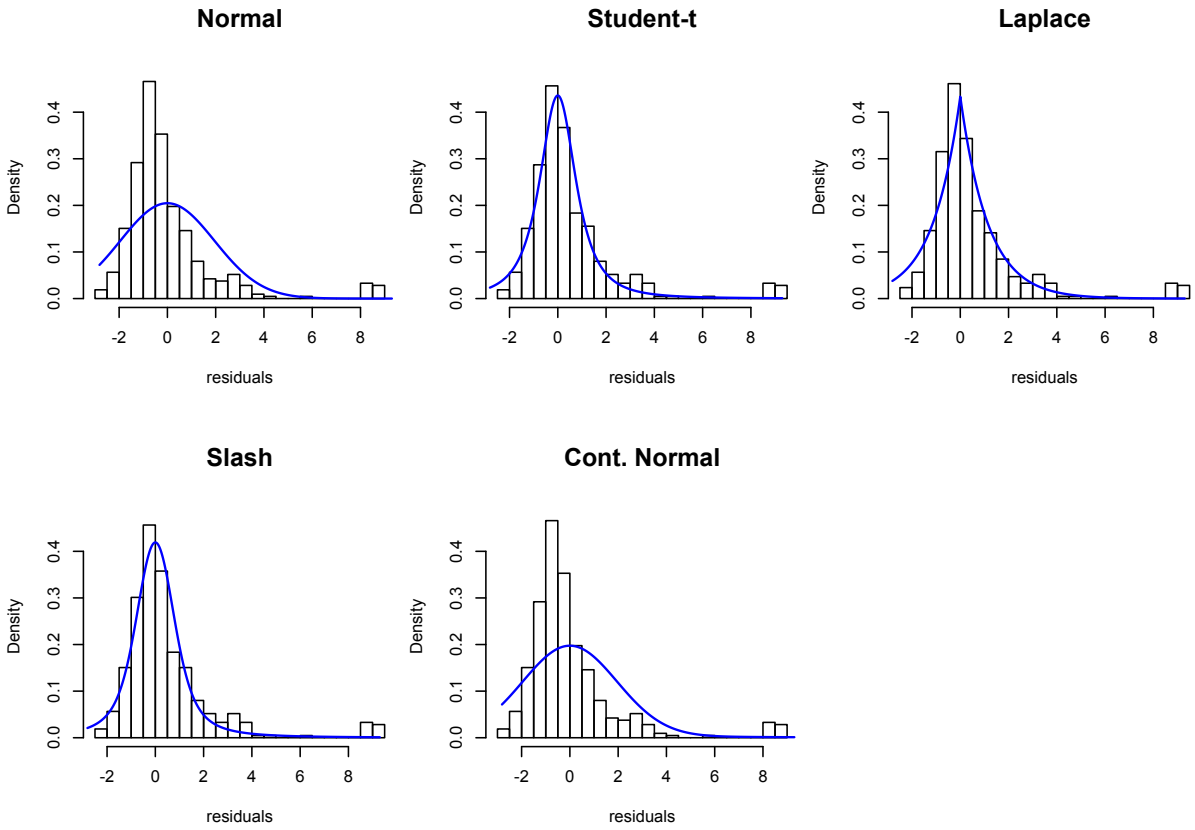

Figure 10: Histograms of the residuals and fitted SKD densities for the median quantile regression model in Subsection 4.1. Plots provided by call the `R Log.lqr` function.

```
Call:
best.lqr(y=score,x=cbind(1,type),p=0.5,criterion = "AIC")

-------------------------------------------------------------
Logistic Quantile Linear Regression using SKD family
-------------------------------------------------------------
Criterion = AIC
Best fit = Student's $t$
Quantile = 0.5

--------------------------------
Model Likelihood-Based criterion
--------------------------------
Normal  Student T  Laplace    Slash   C. Normal
AIC     2368.638 1517.5694 1566.6484 1518.0514 1690.6018
BIC     2380.795 1529.7257 1578.8047 1530.2077 1702.7580
HQ      2373.441 1522.3718 1571.4509 1522.8538 1695.4042
loglik -1181.319 -755.7847 -780.3242 -756.0257 -842.3009

---------
Estimates
---------
Estimate Std. Error z value  Pr(>|z|)
beta 1 -1.63802    0.30818 -5.31508   0.0000 ***
beta 2  0.63249    0.21659  2.92023   0.0035  **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

sigma = 0.81134
nu    = 2.00005
```

Figure 11: An example of outputs from `R lqr` package for median quantile regression model in Subsection 4.1.