Qtools: A Collection of Models and Tools for Quantile Inference

by Marco Geraci

Abstract Quantiles play a fundamental role in statistics. The quantile function defines the distribution of a random variable and, thus, provides a way to describe the data that is specular but equivalent to that given by the corresponding cumulative distribution function. There are many advantages in working with quantiles, starting from their properties. The renewed interest in their usage seen in the last years is due to the theoretical, methodological, and software contributions that have broadened their applicability. This paper presents the R package **Qtools**, a collection of utilities for unconditional and conditional quantiles.

Introduction

Quantiles have a long history in applied statistics, especially the median. The analysis of astronomical data by Galileo Galilei in 1632 (Hald, 2003, p.149) and geodic measurements by Roger Boscovich in 1757 (Koenker, 2005, p.2) are presumably the earliest examples of application of the least absolute deviation (L_1) estimator in its, respectively, unconditional and conditional forms. The theoretical studies on quantiles of continuous random variables started to appear in the statistical literature of the 20th century. In the case of discrete data, studies have somewhat lagged behind most probably because of the analytical drawbacks surrounding the discontinuities that characterise discrete quantile functions. Some forms of approximation to continuity have been recently proposed to study the large sample behavior of quantile estimators. For example, Ma et al. (2011) have demonstrated the asymptotic normality of unconditional sample quantiles based on the definition of the mid-distribution function (Parzen, 2004). Machado and Santos Silva (2005) proposed inferential approaches to the estimation of conditional quantiles for counts based on data jittering.

Functions implementing quantile methods can be found in common statistical software. A considerable number of R packages that provide such functions are available on the Comprehensive R Archive Network (CRAN). The base package **stats** contains basic functions to estimate sample quantiles or compute quantiles of common parametric distributions. The **quantreg** package (Koenker, 2013) is arguably a benchmark for distribution-free estimation of linear quantile regression models, as well as the base for other packages which make use of linear programming (LP) algorithms (Koenker and D'Orey, 1987; Koenker and Park, 1996). Other contributions to the modelling of conditional quantile functions include packages for Bayesian regression, e.g. **bayesQR** (Benoit et al., 2014) and **BSquare** (Smith and Reich, 2013), and the **lqmm** package (Geraci and Bottai, 2014; Geraci, 2014) for random-effects regression.

The focus of this paper is on the R package **Qtools**, a collection of models and tools for quantile inference. These include commands for

- quantile-based analysis of the location, scale and shape of a distribution;
- transformation-based quantile regression;
- goodness of fit and restricted quantile regression;
- quantile regression for discrete data;
- quantile-based multiple imputation.

The emphasis will be put on the first two topics listed above as they represent the main contribution of the package, while a short description of the other topics is given for completeness.

Unconditional quantiles

Definition and estimation of quantiles

Let Y be a random variable with cumulative distribution function (CDF) F_Y and support S_Y. The CDF calculated at $y \in S_Y$ returns the probability $F_Y(y) \equiv p = \Pr(Y \leq y)$. The quantile function (QF) is defined as $Q(p) = \inf_y\{F_Y(y) \geq p\}$, for $0 < p < 1$. (Some authors consider $0 \leq p \leq 1$. For practical purposes, it is simpler to exclude the endpoints 0 and 1.) When F_Y is continuous and strictly monotone (hence, $f_Y(y) \equiv F'_Y(y) > 0$ for all $y \in S_Y$), the quantile function is simply the inverse of F_Y. In other cases, the quantile p is defined, by convention, as the smallest value y such that $F_Y(y)$ is at least p.

Quantiles enjoy a number of properties. An excellent overview is given by Gilchrist (2000). In particular, the **Q-tranformation rule** (Gilchrist, 2000) or equivariance to monotone transformations states that if $h(\cdot)$ is a non-decreasing function on $\mathbb{R}$, then $Q_{h(Y)}(p) = h\{Q_Y(p)\}$. Hence $Q_Y(p) = h^{-1}\left\{Q_{h(Y)}(p)\right\}$. Note that this property does not generally hold for the expected value.

Sample quantiles for a random variable Y can be calculated in a number of ways, depending on how they are defined (Hyndman and Fan, 1996). For example, the function `quantile()` in the base package **stats** provides nine different sample quantile estimators, which are based on the sample order statistics or the inverse of the empirical CDF. These estimators are distribution-free as they do not depend on any parametric assumption about F (or Q).

Let $Y_1, Y_2, \ldots, Y_n$ be a sample of n independent and identically distributed (iid) observations from the population F_Y. Let ξ_p denote the pth population quantile and $\hat{\xi}_p$ the corresponding sample quantile. (The subscripts will be dropped occasionally to ease notation, e.g. F will be used in place of F_Y or ξ in place of ξ_p.) In the continuous case, it is well known that $\sqrt{n}\left(\hat{\xi}_p - \xi_p\right)$ is approximately normal with mean zero and variance

$$\omega^2 = \frac{p(1-p)}{\{f_Y(\xi_p)\}^2}. \tag{1}$$

A more general result is obtained when the Y_i's, $i = 1, \ldots, n$, are independent but not identically distributed (*nid*). The density evaluated at the pth quantile, $f(\xi_p)$, is called the *density-quantile function* by Parzen (1979). Its reciprocal, $s(p) \equiv 1/f(\xi_p)$, is called the *sparsity function* (Tukey, 1965) or *quantile-density function* (Parzen, 1979).

As mentioned previously, the discontinuities of F_Y when Y is discrete represent a mathematical inconvenience. Ma et al. (2011) derived the asymptotic distribution of the sample mid-quantiles, that is, the sample quantiles based on the mid-distribution function (mid-CDF). The latter is defined as $F_Y^{mid}(y) = F_Y(y) - 0.5p_Y(y)$, where $p_Y(y)$ denotes the probability mass function (Parzen, 2004). In particular, they showed that, as n becomes large, $\sqrt{n}\left(\hat{\xi}_p^{mid} - \xi_p\right)$ is approximately normal with mean 0. Under iid assumptions, the expression for the sampling variance is similar to that in (1); see Ma et al. (2011) for details.

The package **Qtools** provides the functions `midecdf()` and `midquantile()`, which return objects of class `"midecdf"` or `"midquantile"`, respectively, containing: the values or the probabilities at which mid-cumulative probabilities or mid-quantiles are calculated (`x`), the mid-cumulative probabilities or the mid-quantiles (`y`), and the functions that linearly interpolate those coordinates (`fn`). An example is shown below using data simulated from a Poisson distribution.

```
> library("Qtools")
> set.seed(467)
> y <- rpois(1000, 4)
> pmid <- midecdf(y)
> xmid <- midquantile(y, probs = pmid$y)
> pmid

Empirical mid-ECDF
Call:
midecdf(x = y)

> xmid

Empirical mid-ECDF
Call:
midquantile(x = y, probs = pmid$y)
```

A confidence interval for sample mid-quantiles can be obtained using `confint.midquantile()`. This function is applied to the output of `midquantile()` and returns an object of class `"data.frame"` containing sample mid-quantiles, lower and upper bounds of the confidence intervals of a given level (95% by default), along with standard errors as an attribute named `stderr`. This is shown below using the sample y generated in the previous example.

```
> xmid <- midquantile(y, probs = 1:3/4)
> x <- confint(xmid, level = 0.95)
> x

    midquantile    lower    upper
```

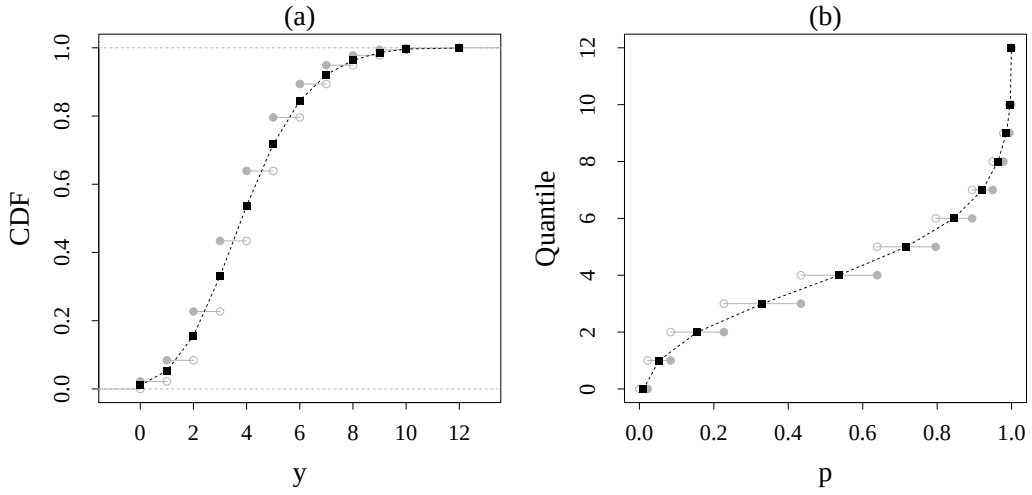

Figure 1: Cumulative distribution (a) and quantile (b) functions for simulated Poisson data. The ordinary cumulative distribution function (CDF) and quantile function (QF) are represented by step-functions (grey lines), with the convention that, at the point of discontinuity or *jump*, the function takes its value corresponding to the ordinate of the filled circle as opposed to that of the hollow circle. The mid-CDF and mid-QF are represented by filled squares, while the piecewise linear functions (dashed lines) connecting the squares represent continuous versions of, respectively, the ordinary CDF and QF.

```
25%    2.540000 2.416462 2.663538
50%    3.822816 3.693724 3.951907
75%    5.254902 5.072858 5.436946

> attr(x, "stderr")
[1] 0.06295447 0.06578432 0.09276875
```

Finally, a plot method is available for both `midecdf()` and `midquantile()` objects. An illustration is given in Figure 1. The mid-distribution and mid-quantile functions are discrete and their values are marked by filled squares. The piecewise linear functions connecting the filled squares represent continuous versions of the CDF and QF which interpolate between the steps of, respectively, the ordinary CDF and quantile functions. Note that the argument `jumps` is a logical value indicating whether values at jumps should be marked.

```
> par(mfrow = c(1,2))
> plot(pmid, xlab = "y", ylab = "CDF", jumps = TRUE)
> points(pmid$x, pmid$y, pch = 15)
> plot(xmid, xlab = "p", ylab = "Quantile", jumps = TRUE)
> points(xmid$x, xmid$y, pch = 15)
```

LSS - Location, scale and shape of a distribution

Since the cumulative distribution and quantile functions are two sides of the same coin, the location, scale, and shape (LSS) of a distribution can be examined using one or the other. Well-known quantile-based measures of location and scale are the median and inter-quartile range (IQR), respectively. Similarly, there are also a number of quantile-based measures for skewness and kurtosis (Groeneveld and Meeden, 1984; Groeneveld, 1998; Jones et al., 2011).

Define the *central* portion of the distribution as that delimited by the quantiles $Q(p)$ and $Q(1-p)$, $0 < p < 0.5$, and define the *tail* portion as that lying outside these quantiles. Let $IPR(p) = Q(1-p) - Q(p)$ denote the inter-quantile range at level p. Building on the results by Horn (1983) and Ruppert (1987), Staudte (2014) considered the following identity:

$$\underbrace{\frac{IPR(p)}{IPR(r)}}_{\text{kurtosis}} = \underbrace{\frac{IPR(p)}{IPR(q)}}_{\text{tail-weight}} \cdot \underbrace{\frac{IPR(q)}{IPR(r)}}_{\text{peakedness}}, \tag{2}$$

where $0 < p < q < r < 0.5$. These quantile-based measures of shape are sign, location and scale invariant. As compared to moment-based indices, they are also more robust to outliers and easier to interpret (Groeneveld, 1998; Jones et al., 2011).

It is easy to verify that a quantile function can be written as

$$Q(p) = \underbrace{Q(0.5)}_{\text{median}} + \frac{1}{2}\underbrace{IPR(0.25)}_{\text{IQR}} \cdot \underbrace{\frac{IPR(p)}{IPR(0.25)}}_{\text{shape index}} \cdot \Bigg(\underbrace{\frac{Q(p) + Q(1-p) - 2Q(0.5)}{IPR(p)}}_{\text{skewness index}} - 1 \Bigg). \tag{3}$$

This identity establishes a relationship between the location (median), scale (IQR) and shape of a distribution. (This identity appears in Gilchrist (2000, p.74) with an error of sign. See also Benjamini and Krieger (1996, eq.1).) The quantity $IPR(p)/IPR(0.25)$ in (3) is loosely defined as the *shape index* (Gilchrist, 2000, p.72), although it can be seen as the tail-weight measure given in (2) when $p < 0.25$. For symmetric distributions, the contribution of the skewness index vanishes. Note that the skewness index not only is location and scale invariant, but is also bounded between -1 and 1 (as opposed to the Pearson's third standardised moment which can be infinite or even undefined). When this index is near the bounds -1 or 1, then $Q(1-p) \approx Q(0.5)$ or $Q(p) \approx Q(0.5)$, respectively.

The function `qlss()` provides a quantile-based LSS summary with the indices defined in (3) of either a theoretical or an empirical distribution. It returns an object of class `"qlss"`, which is a list containing measures of location (median), scale (IQR and IPR), and shape (skewness and shape indices) for each of the probabilities specified in the argument `probs` (by default, `probs` = 0.1). The quantile-based LSS summary of the normal distribution is given in the example below for $p = 0.1$. The argument `fun` can take any quantile function whose probability argument is named 'p' (this is the case for many standard quantile functions in R, e.g. `qt()`, `qchisq()`, `qf()`, etc.).

```
> qlss(fun = "qnorm", probs = 0.1)

call:
qlss.default(fun = "qnorm", probs = 0.1)

Unconditional Quantile-Based Location, Scale, and Shape

** Location **
Median
[1] 0
** Scale **
Inter-quartile range (IQR)
[1] 1.34898
Inter-quantile range (IPR)
     0.1
2.563103
** Shape **
Skewness index
0.1
  0
Shape index
     0.1
1.900031
```

An empirical example is now illustrated using the `faithful` data set, which contains 272 observations on waiting time (minutes) between eruptions and the duration (minutes) of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Summary statistics are given in Table 1.

	Minimum	Q1	Q2	Q3	Maximum
Waiting time	43.0	58.0	76.0	82.0	96.0
Duration	1.6	2.2	4.0	4.5	5.1

Table 1: Minimum, maximum and three quartiles (Q1, Q2, Q3) for waiting time and duration in the Old Faithful Geyser data set.

Suppose the interest is in describing the distribution of waiting times. The density is plotted in Figure 2, along with the mid-quantile function. The distribution is bimodal with peaks at around 54

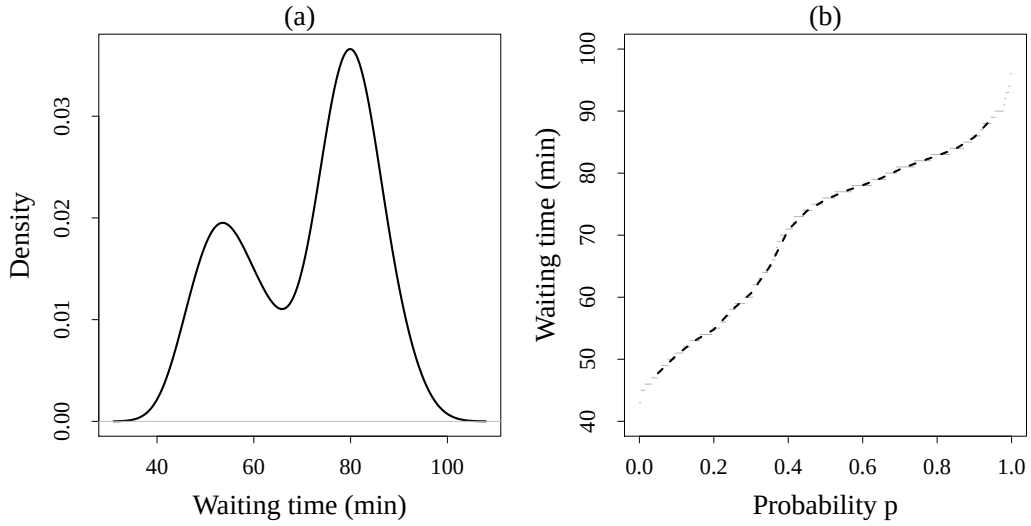

Figure 2: Estimated density (a) and empirical mid-quantile (b) functions of waiting time between eruptions in the Old Faithful Geyser data set.

and 80 minutes. Note that the arguments of the base function `quantile()`, including the argument `type`, can be passed on to `qlss()`.

```
> y <- faithful$waiting
> par(mfrow = c(1,2))
> plot(density(y))
> plot(midquantile(y, probs = p), jumps = FALSE)
> qlss(y, probs = c(0.05, 0.1, 0.25), type = 7)

call:
qlss.numeric(x = y, probs = c(0.05, 0.1, 0.25), type = 7)

Unconditional Quantile-Based Location, Scale, and Shape

** Location **
Median
[1] 76
** Scale **
Inter-quartile range (IQR)
[1] 24
Inter-quantile range (IPR)
0.05  0.1 0.25
  41   35   24
** Shape **
Skewness index
      0.05        0.1       0.25
-0.3658537 -0.4285714 -0.5000000
Shape index
    0.05      0.1     0.25
1.708333 1.458333 1.000000
```

At $p = 0.1$, the skewness index is approximately -0.43, which denotes a rather strong left asymmetry. As for the shape index, which is equal to 1.46, one could say that the tails of this distribution weigh less than those of a normal distribution (1.90), though of course a comparison between unimodal and bimodal distributions is not meaningful.

Conditional quantiles

Linear models

In general, the pth linear QR model is of the form

$$Q_{Y|X}(p) = \mathbf{x}^\top \boldsymbol{\beta}(p) \tag{4}$$

where $\mathbf{x}$ is a k-dimensional vector of covariates (including 1 as first element) and $\boldsymbol{\beta}(p) = [\beta_0(p), \beta_1(p), \ldots, \beta_{k-1}(p)]^\top$ is a vector of coefficients. The *slopes* $\beta_j(p), j = 1, \ldots, k-1$, have the usual interpretation of partial derivatives. For example, in case of the simple model $Q_{Y|X}(p) = \beta_0(p) + \beta_1(p)x$, one obtains

$$\frac{\partial Q_{Y|X}(p)}{\partial x} = \beta_1(p).$$

If x is a dummy variable, then $\beta_1(p) = Q_{Y|X=1}(p) - Q_{Y|X=0}(p)$, i.e. the so-called *quantile treatment effect* (Doksum, 1974; Lehmann, 1975; Koenker and Xiao, 2002). Estimation can be carried out using LP algorithms which, given a sample $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, solve

$$\min_{b \in \mathbb{R}^k} \sum_{i=1}^{n} \kappa_p \left(y_i - \mathbf{x}_i^\top \mathbf{b} \right),$$

where $\kappa_p(u) = u(p - I(u < 0))$, $0 < p < 1$, is the check loss function. Large-n approximation of standard errors can be obtained from the sampling distribution of the linear quantile estimators (Koenker and Bassett, 1978).

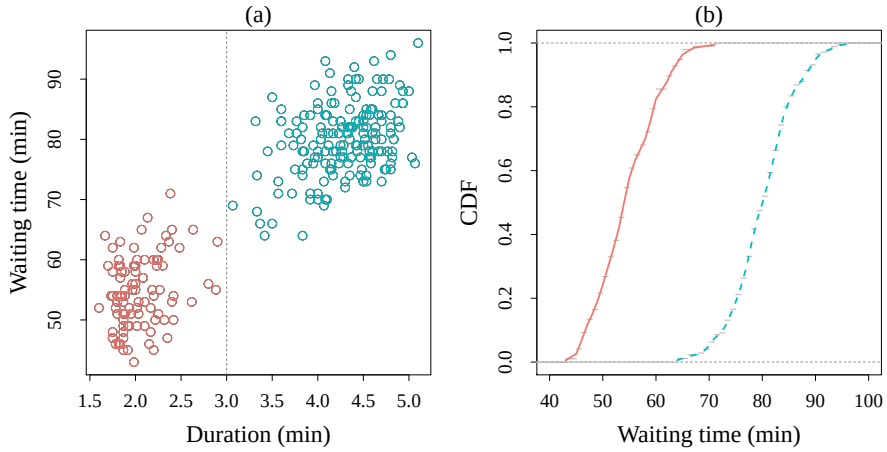

Figure 3: (a) Waiting times between eruptions against durations of eruptions (dashed vertical line drawn at 3 minutes) in the Old Faithful Geyser data set. (b) Mid-CDF of waiting time by duration of eruption (solid line, shorter than 3 minutes; dashed line, longer than 3 minutes).

Waiting times between eruptions are plotted against the durations of the eruptions in Figure 3. Two clusters of observations can be defined for durations below and above 3 minutes (see also Azzalini and Bowman, 1990). The distribution shows a strong bimodality as already illustrated in Figure 2. A dummy variable for durations equal to or longer than 3 minutes is created to define the two distributions and included as covariate X in a model as the one specified in (4). The latter is then fitted to the Old Faithful Geyser data using the function `rq()` in the package **quantreg** for $p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

```
> require("quantreg")
> y <- faithful$waiting
> x <- as.numeric(faithful$eruptions >= 3)
> fit <- rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9))
> fit


Call:
rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9))
```

```
Coefficients:
            tau= 0.10 tau= 0.25 tau= 0.50 tau= 0.75 tau= 0.90
(Intercept)        47        50        54        59        63
x                  26        26        26        25        25

Degrees of freedom: 272 total; 270 residual
```

From the output above, it is quite evident that the distribution of waiting times is shifted by an approximately constant amount at all considered values of p. The location-shift hypothesis can be tested by using the Khmaladze test. The null hypothesis is that two distributions, say F_0 and F_1, differ by a pure location shift (Koenker and Xiao, 2002), that is

$$H_0 : F_1^{-1}(p) = F_0^{-1}(p) + \delta_0,$$

where δ_0 is the quantile treatment effect, constant over p. The location–scale-shift specification of the test considers

$$H_0 : F_1^{-1}(p) = \delta_1 F_0^{-1}(p) + \delta_0.$$

The alternative hypothesis is that the model is more complex than the one specified in the null hypothesis. The Khmaladze test is implemented in **quantreg** (see `?quantreg::KhmaladzeTest` for further details). The critical values of the test and corresponding significance levels (Koenker, 2005) are not readily available in the same package. These have been hardcoded in the **Qtools** function `KhmaladzeFormat()` which can be applied to `"KhmaladzeTest"` objects. For the Old Faithful Geyser data, the result of the test is not statistically significant at the 10% level.

```
> kt <- KhmaladzeTest(formula = y ~ x, taus = seq(.05, .95, by = .01),
> KhmaladzeFormat(kt, 0.05)

Khmaladze test for the location-shift hypothesis
Joint test is not significant at 10% level
Test(s) for individual slopes:
 not significant at 10% level
```

Goodness of fit

Distribution-free quantile regression does not require introducing an assumption on the functional form of the error distribution (Koenker and Bassett, 1978), but only weaker quantile restrictions (Powell, 1994). Comparatively, the linear specification of the conditional quantile function in Equation 4 is a much stronger assumption and thus plays an important role for inferential purposes.

The problem of assessing the goodness of fit (GOF) is rather neglected in applications of QR. Although some approaches to GOF have been proposed (Zheng, 1998; Koenker and Machado, 1999; He and Zhu, 2003; Khmaladze and Koul, 2004), there is currently a shortage of software code available to users. The function `GOFTest()` implements a test based on the cusum process of the gradient vector (He and Zhu, 2003). Briefly, the test statistic is given by the largest eigenvalue of

$$n^{-1} \sum_{i}^{n} \mathbf{R}_n(\mathbf{x}_i)\mathbf{R}_n^{\top}(\mathbf{x}_i)$$

where $\mathbf{R}_n(\mathbf{t}) = n^{-1/2} \sum_{j=1}^{n} \psi_p(r_j)\mathbf{x}_j I(\mathbf{x}_j \leq \mathbf{t})$ is the residual cusum (RC) process and $\psi_p(r_j)$ is the derivative of the loss function κ_p calculated for residual $r_j = y_j - \mathbf{x}_j^{\top}\boldsymbol{\beta}(p)$. The sampling distribution of this test statistic is non-normal (He and Zhu, 2003) and a resampling approach is used to obtain the p-value under the null hypothesis.

An example is provided further below using the New York Air Quality data set, which contains 111 complete observations on daily mean ozone (parts per billion – ppb) and solar radiation (Langleys – Ly). For simplicity, wind speed and maximum daily temperature, also included in the data set, are not analysed here.

Suppose that the model of interest is

$$Q_{\text{ozone}}(p) = \beta_0(p) + \beta_1(p) \cdot \text{Solar.R.} \tag{5}$$

Three conditional quantiles ($p \in \{0.1, 0.5, 0.9\}$) are estimated and plotted using the following code:

```
> dd <- airquality[complete.cases(airquality), ]
> dd <- dd[order(dd$Solar.R), ]
> fit.rq <- rq(Ozone ~ Solar.R, tau = c(.1, .5, .9), data = dd)
```

```
> x <- seq(min(dd$Solar.R), max(dd$Solar.R), length = 200)
> yhat <- predict(fit.rq, newdata = data.frame(Solar.R = x))
> plot(Ozone ~ Solar.R, data = dd)
> apply(yhat, 2, function(y, x) lines(x, y), x = x)
```

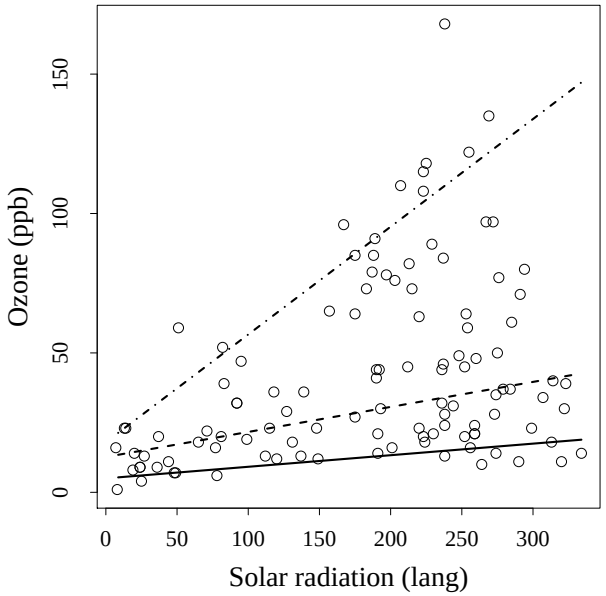

Figure 4: Predicted 10th (solid line), 50th (dashed line), and 90th (dot-dashed line) centiles of ozone conditional on solar radiation in the New York Air Quality data set.

As a function of solar radiation, the median of the ozone daily averages increases by 0.09 ppb for each Ly increase in solar radiation (Figure 4). The 90th centile of conditional ozone shows a steeper slope at 0.39 ppb/Ly, about nine times larger than the slope of the conditional 10th centile at 0.04 ppb/Ly.

The RC test applied to the the object `fit.rq` provides evidence of lack of fit for all quantiles considered, particularly for $p = 0.1$ and $p = 0.5$. Therefore the straight-line model in Equation 5 for these three conditional quantiles does not seem to be appropriate. The New York Air Quality data set will be analysed again in the next section, where a transformation-based approach to nonlinear modelling is discussed.

```
> gof.rq <- GOFTest(fit.rq, alpha = 0.05, B = 1000, seed = 987)
> gof.rq

Goodness-of-fit test for quantile regression based on the cusum process
Quantile 0.1: Test statistic = 0.1057; p-value = 0.001
Quantile 0.5: Test statistic = 0.2191; p-value = 0
Quantile 0.9: Test statistic = 0.0457; p-value = 0.018
```

Transformation models

Complex dynamics may result in nonlinear effects in the relationship between the covariates and the response variable. For instance, in kinesiology, pharmacokinetics, and enzyme kinetics, the study of the dynamics of an agent in a system involves the estimation of nonlinear models; phenomena like human growth, certain disease mechanisms and the effects of harmful environmental substances such as lead and mercury, may show strong nonlinearities over time. In this section, the linear model is abandoned in favor of a more general model of the type

$$Q_{Y|X}(p) = g\left\{\mathbf{x}^\top \boldsymbol{\beta}(p)\right\}, \tag{6}$$

for some real-valued function g. If g is nonlinear, the alternative approaches to conditional quantile modelling are

Nonlinear parametric models which may provide substantive interpretability, possibly parsimonious (in general more parsimonious than polynomials), and valid beyond the observed range

of the data. A nonlinear model depends on either prior knowledge of the phenomenon or the introduction of new, strong theory to explain the observed relationship with potential predictive power. Estimation may present challenges;

Polynomial models and smoothing splines falling under the label of *nonparametric regression*, in which the complexity of the model is approximated by a sequence of locally linear polynomials (a naïve global polynomial trend can be considered to be a special case). A nonparametric model need not introducing strong assumptions about the relationship and is essentially data-driven. Estimation is based on linear approximations and, typically, requires the introduction of a penalty term to control the degree of smoothing; and

Transformation models a flexible, parsimonious family of parametric transformations is applied to the response seeking to obtain approximate linearity on the transformed scale. The data provide information about the "best" transformation among a family of transformations. Estimation is facilitated by the application of methods for linear models.

The focus of this section is on the third approach. More specifically the functions available in **Qtools** refer to the methods for transformation-based QR models developed by Powell (1991), Chamberlain (1994), Mu and He (2007), Dehbi et al. (2016) and Geraci and Jones (2015). Examples of approaches to nonlinear QR based on parametric models or splines can be found in Koenker and Park (1996) and Yu and Jones (1998), respectively.

The goal of the transformation-based QR is to fit the model

$$Q_{h\left(Y;\lambda_p\right)}(p) = \mathbf{x}^\top \boldsymbol{\beta}(p). \tag{7}$$

The assumption is that the transformation h is the inverse of g, $h\left(Y;\lambda_p\right) \equiv g^{-1}(Y)$, so that the pth quantile function of the transformed response variable is linear. (In practice, it is satisfactory to achieve approximate linearity.) The parameter λ_p is a low-dimensional parameter that gives some flexibility to the shape of the transformation and is estimated from the data. In general, the interest is on predicting $Q_{Y|X}(p)$ and estimating the effects of the covariates on $Q_{Y|X}(p)$. If h is a non-decreasing function on $\mathbb{R}$ (as is the case for all transformations considered here), predictions can be easily obtained from (7) by virtue of the equivariance property of quantiles,

$$Q_{Y|X}(p) = h^{-1}\left\{\mathbf{x}^\top \boldsymbol{\beta}(p); \lambda_p\right\}. \tag{8}$$

The marginal effect of the jth covariate x_j can be obtained by differentiating the quantile function $Q_{Y|X}(p)$ with respect to x_j. This can be written as the derivative of the composition $Q \circ \eta$, i.e.

$$\frac{\partial Q(p)}{\partial x_j} = \frac{\partial Q(p)}{\partial \eta(p)} \cdot \frac{\partial \eta(p)}{\partial x_j}, \tag{9}$$

$\eta(p) = \mathbf{x}^\top \boldsymbol{\beta}(p)$. Once the estimates $\hat{\boldsymbol{\beta}}(p)$ and $\hat{\lambda}_p$ are obtained, these can be plugged in Equations 8 and 9.

The package **Qtools** provides several transformation families, namely the Box–Cox (Box and Cox, 1964), Aranda-Ordaz (Aranda-Ordaz, 1981), and Jones (Jones, 2007; Geraci and Jones, 2015) transformations. A distinction between these families is made in terms of the support of the response variable to which the transformation is applied and the number of transformation parameters. The Box–Cox model is a one-parameter family of transformations which applies to singly bounded variables, $y > 0$. The Aranda-Ordaz symmetric and asymmetric transformations too have one parameter and are used when responses are bounded on the unit interval, $0 < y < 1$ (doubly bounded). Geraci and Jones (2015) developed two families of transformations which can be applied to either singly or doubly bounded responses:

Proposal I transformations with one parameter and assuming both symmetric and asymmetric forms;

Proposal II transformations with two parameters, with one parameter modelling the symmetry (or lack thereof) of the transformation.

Originally, Box and Cox (1964) proposed using power transformations to address lack of linearity, homoscedasticity and normality of the residuals in mean regression modelling. Sakia (1992, p.175) reported that "seldom does this transformation fulfil the basic assumptions of linearity, normality and homoscedasticity simultaneously as originally suggested by Box & Cox (1964). The Box-Cox transformation has found more practical utility in the empirical determination of functional relationships in a variety of fields, especially in econometrics".

Indeed, the practical utility of power transformations has been long recognised in QR modelling (Powell, 1991; Buchinsky, 1995; Chamberlain, 1994; Mu and He, 2007). Model 7 is the Box–Cox QR

model if

$$h_{BC}\left(Y;\lambda_p\right) = \begin{cases} \dfrac{Y^{\lambda_p}-1}{\lambda_p} & \text{if } \lambda_p \neq 0 \\ \log Y & \text{if } \lambda_p = 0. \end{cases} \tag{10}$$

Note that when $\lambda_p \neq 0$, the range of this transformation is not $\mathbb{R}$ but the singly bounded interval $\left(-1/\lambda_p, \infty\right)$. This implies that the inversion in (8) is defined only for $\lambda_p \mathbf{x}^\top \boldsymbol{\beta}(p) + 1 > 0$.

The *symmetric Aranda-Ordaz transformation* is given by

$$h_{AOs}\left(Y;\lambda_p\right) = \begin{cases} \dfrac{2}{\lambda_p} \dfrac{Y^{\lambda_p}-(1-Y)^{\lambda_p}}{Y^{\lambda_p}+(1-Y)^{\lambda_p}} & \text{if } \lambda_p \neq 0, \\ \log\left(\dfrac{Y}{1-Y}\right) & \text{if } \lambda_p = 0. \end{cases} \tag{11}$$

(The symmetry here is that $h_{AOs}\left(\theta;\lambda_p\right) = -h_{AOs}\left(1-\theta;\lambda_p\right) = h_{AOs}\left(\theta;-\lambda_p\right)$.) There is a range problem with this transformation too since, for all $\lambda_p \neq 0$, the range of h_{AOs} is not $\mathbb{R}$, but $\left(-2/|\lambda_p|, 2/|\lambda_p|\right)$. The *asymmetric Aranda-Ordaz transformation* is given by

$$h_{AOa}\left(Y;\lambda_p\right) = \begin{cases} \log\left\{\dfrac{(1-Y)^{-\lambda_p}-1}{\lambda_p}\right\} & \text{if } \lambda_p \neq 0, \\ \log\left\{-\log\left(1-Y\right)\right\} & \text{if } \lambda_p = 0. \end{cases} \tag{12}$$

For $\lambda_p = 0$, this is equivalent to the complementary log-log. The asymmetric Aranda-Ordaz transformation does have range $\mathbb{R}$. Note that $h_{AOa}(Y;1) = \log(Y/(1-Y))$, i.e. the transformation is symmetric.

To overcome range problems, which give rise to computational difficulties, Geraci and Jones (2015) proposed to use instead one-parameter transformations with range $\mathbb{R}$. Proposal I is written in terms of the variable (say) W, where

$$h_I\left(W;\lambda_p\right) = \begin{cases} \dfrac{1}{2\lambda_p}\left(W^{\lambda_p} - \dfrac{1}{W^{\lambda_p}}\right) & \text{if } \lambda_p \neq 0 \\ \log W & \text{if } \lambda_p = 0, \end{cases} \tag{13}$$

which takes on four forms depending on the relationship of W to Y, as described in Table 2. For each of domains $(0,\infty)$ and $(0,1)$, there are symmetric and asymmetric forms.

Support of Y	Symmetric	Asymmetric
$(0,\infty)$	$W = Y$ $h_{Is}\left(Y;\lambda_p\right)$	$W = \log(1+Y)$ $h_{Ia}\left(Y;\lambda_p\right)$
$(0,1)$	$W = Y/(1-Y)$ $h_{Is}\left(Y;\lambda_p\right)$	$W = -\log(1-Y)$ $h_{Ia}\left(Y;\lambda_p\right)$

Table 2: Choices of W and corresponding notation for transformations based on (13).

Since the transformation in (13) has range $\mathbb{R}$ for all λ_p, it admits an explicit inverse transformation. In addition, in the case of a single covariate, every estimated quantile that results will be monotone increasing, decreasing or constant, although different estimated quantiles can have different shapes from this collection. Geraci and Jones (2015) also proposed a transformation that unifies the symmetric and asymmetric versions of h_I into a single two-parameter transformation, namely

$$h_{II}\left(W;\lambda_p\right) = h_I\left(W_{\delta_p};\lambda_p\right), \tag{14}$$

where h_I is given in (13) and

$$W_{\delta_p} = h_{BC}\left(1+W;\delta_p\right) = \begin{cases} \dfrac{(1+W)^{\delta_p}-1}{\delta_p} & \text{if } \delta_p > 0 \\ \log(1+W) & \text{if } \delta_p = 0, \end{cases}$$

with $W = Y$, if $Y > 0$, and $W = Y/(1-Y)$, if $Y \in (0,1)$. The additional parameter δ_p controls the asymmetry: symmetric forms of h_I correspond to $\delta_p = 1$ while asymmetric forms of h_I to $\delta_p = 0$.

All transformation models discussed above can be fitted using a two-stage (TS) estimator (Chamberlain, 1994; Buchinsky, 1995) whereby $\boldsymbol{\beta}(p)$ is estimated conditionally on a fine grid of values for the transformation parameter(s). Alternatively, point estimation can be approached using the RC process (Mu and He, 2007), which is akin to the process that leads to the RC test introduced in the previous section. The RC estimator avoids the troublesome inversion of the Box-Cox and Aranda-Ordaz transformations, but it is computationally more intensive than the TS estimator.

There are several methods for interval estimation, including those based on large-n approximations and the ubiquitous bootstrap. Both the TS and RC estimators have an asymptotic normal distribution. The large-sample properties of the TS estimator for monotonic quantile regression models have been studied by Powell (1991) (see also Chamberlain, 1994; Machado and Mata, 2000). Under regularity conditions, it can be shown that the TS estimator is unbiased and will converge to a normal distribution with a sandwich-type limiting covariance matrix which is easy to calculate. In contrast, the form of the covariance matrix of the sampling distribution for the RC estimator is rather complicated and its estimation requires resampling (Mu and He, 2007). Finally, if the transformation parameter is assumed to be known, then conditional inference is apposite. In this case, the estimation procedures simplify to those for standard quantile regression problems.

In **Qtools**, model fitting for one-parameter transformation models can be carried out using the function `tsrq()`. The `formula` argument specifies the model for the linear predictor as in (7), while the argument `tsf` provides the desired transformation h as specified in Equations 10-13: `"bc"` for the Box–Cox model, `"ao"` for Aranda-Ordaz families, and `"mcjI"` for proposal I transformations. Additional arguments in the function `tsrq()` include

`symmetry` a logical flag to specify the symmetric or asymmetric version of `"ao"` and `"mcjI"`;

`dbounded` a logical flag to specify whether the response variable is doubly bounded (default is strictly positive, i.e. singly bounded);

`lambda` a numerical vector to define the grid of values for estimating λ_p; and `conditional`, a logical flag indicating whether λ_p is assumed to be known (in which case the argument `lambda` provides such known value).

There are other functions to fit transformation models. The function `rcrq()` fits one-parameter transformation models using the RC estimator. The functions `tsrq2()` and `nlrq2()` are specific to Geraci and Jones's (2015) Proposal II transformations. The former employs a two-way grid search while the latter is based on Nelder-Mead optimization as implemented in `optim()`. Simulation studies in Geraci and Jones (2015) suggest that, although computationally slower, a two grid search is numerically more stable than the derivative-free approach.

A summary of the basic differences between all fitting functions is given in Table 3. The table also shows the available methods in `summary.rqt()` to estimate standard errors and confidence intervals for the model's parameters. Unconditional inference is carried out jointly on $\boldsymbol{\beta}(p)$ and the transformation parameter by means of bootstrap using the package **boot** (Canty and Ripley, 2014; Davison and Hinkley, 1997). Large-n approximations (Powell, 1991; Chamberlain, 1994; Machado and Mata, 2000) are also available for the one-parameter TS estimator under iid or nid assumptions.

When `summary.rqt()` is executed with the argument `conditional = TRUE`, confidence interval estimation for $\boldsymbol{\beta}_p$ is performed with one of the several methods developed for linear quantile regression estimators (Koenker, 2005, p.110) (see options `"rank"`, `"iid"`, `"nid"`, `"ker"`, and `"boot"` in `quantreg::summary.rq()`).

In the New York Air Quality data example, a linear model was found unsuitable to describe the relationship between ozone and solar radiation. At closer inspection, Figure 4 reveals that the conditional distribution of ozone may in fact be nonlinearly associated with solar radiation, at least for some of the conditional quantiles. The model

$$Q_{h_{Is}\{\text{ozone}\}}(p) = \beta_0(p) + \beta_1(p)\cdot \text{Solar.R}, \tag{15}$$

where h_{Is} denotes the symmetric version of (13) for a singly bounded response variable, is fitted for the quantiles $p \in \{0.1, 0.5, 0.9\}$ using the following code:

Function name	Transformation parameters	Estimation	Standard errors or confidence intervals Unconditional	Conditional
`tsrq()`	1	Two-stage	`"iid"`, `"nid"`, `"boot"`	All types
`rcrq()`	1	Residual cusum process	`"boot"`	All types
`tsrq2()`	2	Two-stage	`"boot"`	All types
`nlrq2()`	2	Nelder–Mead	`"boot"`	–

Table 3: Transformation-based quantile regression in package **Qtools**. *All types* consists of options `"rank"`, `"iid"`, `"nid"`, `"ker"`, and `"boot"` as provided by function `summary()` in package **quantreg**.

```
> system.time(fit.rqt <- tsrq(Ozone ~ Solar.R, data = dd, tsf = "mcjI",
+    symm = TRUE, dbounded = FALSE, lambda = seq(1, 3, by = 0.005),
+    conditional = FALSE, tau = c(.1, .5, .9)))
   user  system elapsed
    0.5     0.0     0.5
> fit.rqt

call:
tsrq(formula = Ozone ~ Solar.R, data = dd, tsf = "mcjI", symm = TRUE,
    dbounded = FALSE, lambda = seq(1, 3, by = 0.005), conditional = FALSE,
    tau = c(0.1, 0.5, 0.9))

Proposal I symmetric transformation (singly bounded response)

Optimal transformation parameter:
tau = 0.1 tau = 0.5 tau = 0.9
    2.210     2.475     1.500

Coefficients linear model (transformed scale):
             tau = 0.1  tau = 0.5 tau = 0.9
(Intercept) -3.3357578 -48.737341 16.557327
Solar.R      0.4169697   6.092168  1.443407

Degrees of freedom: 111 total; 109 residual
```

The TS estimator makes a search for λ_p over the grid $1.000, 1.005, \ldots, 2.995, 3.000$. The choice of the search interval usually results from a compromise between accuracy and performance: the coarser the grid, the faster the computation but the less accurate the estimate. A reasonable approach would be to start with a coarse, wide-ranging grid (e.g. `seq(-5,5,by = 0.5)`), then center the interval about the resulting estimate using a finer grid, and re-fit the model.

The output above reports the estimates $\hat{\boldsymbol{\beta}}(p)$ and $\hat{\lambda}_p$ for each quantile level specified in `tau`. Here, the quantities of interest are the predictions on the ozone scale and the marginal effect of solar radiation, which can obtained using the function `predict.rqt()`.

```
> x <- seq(9, 334, length = 200)
> qhat <- predict(fit.rqt, newdata = data.frame(Solar.R = x),
+    type = "response")
> dqhat <- predict(fit.rqt, newdata = data.frame(Solar.R = x),
+    type = "maref", namevec = "Solar.R")
The linear component of the marginal effect is calculated as derivative of
 Ozone ~ beta1 * Solar.R
 with respect to Solar.R
```

The calculations above are based on a sequence of 200 ozone values in the interval $[9, 334]$ Ly, as provided via the argument `newdata` (if this argument is missing, the function returns the fitted values). There are three types of predictions available:

`link` predictions of conditional quantiles on the transformed scale (7), i.e. $\hat{Q}_{h(Y;\hat{\lambda}_p)}(p) = \mathbf{x}^T\hat{\boldsymbol{\beta}}(p)$;

`response` predictions of conditional quantiles on the original scale (8), i.e. $\hat{Q}_{Y|X}(p) = h^{-1}\left\{\mathbf{x}^\top\hat{\boldsymbol{\beta}}(p); \hat{\lambda}_p\right\}$; and

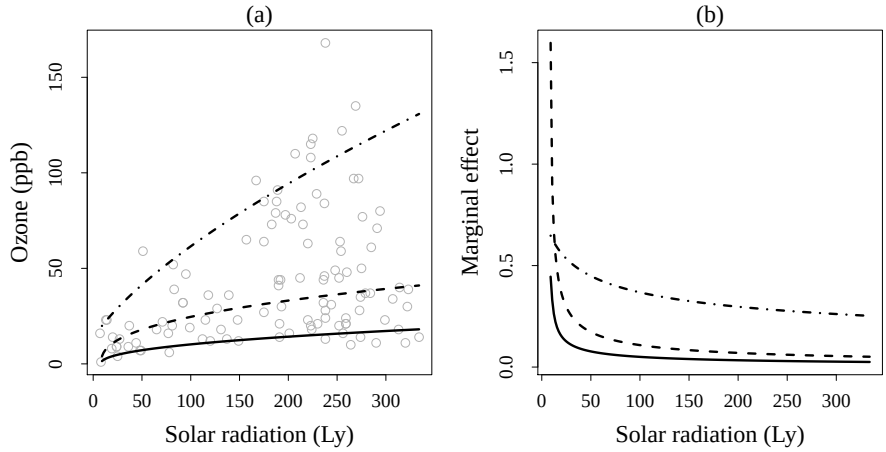

Figure 5: Predicted 10th (solid line), 50th (dashed line), and 90th (dot-dashed line) centiles of ozone conditional on solar radiation (a) and corresponding estimated marginal effects (b) using the symmetric proposal I transformation in the New York Air Quality data set.

`maref` predictions of the marginal effect (9).

In the latter case, the argument `namevec` is used to specify the name of the covariate with respect to which the marginal effect has to be calculated. The function `maref.rqt()` computes derivatives symbolically using the **stats** function `deriv()` and these are subsequently evaluated numerically. While the nonlinear component of the marginal effect in Equation 9 (i.e. $\partial Q(p)/\partial \eta(p)$) is rather straightforward to derive for any of the transformations (10)-(13), the derivative of the linear predictor (i.e. $\partial \eta(p)/\partial x_j$) requires parsing the `formula` argument in order to obtain an expression suitable for `deriv()`. The function `maref.rqt()` can handle simple expressions with common functions like `log()`, `exp()`, etc. , interaction terms, and `"AsIs"` terms (i.e. `I()`). However, using functions that are not recognised by `deriv()` will trigger an error.

The predicted quantiles of ozone and the marginal effects of solar radiation are plotted in Figure 5 using the following code:

```
> par(mfrow = c(1, 2))
> plot(Ozone ~ Solar.R, data = dd, xlab = "Solar radiation (lang)",
+    ylab = "Ozone (ppb)")
> for(i in 1:3) lines(x, qhat[ ,i], lty = c(1, 2, 4)[i], lwd = 2)
> plot(range(x), range(dqhat), type = "n", xlab = "Solar radiation (lang)",
+    ylab = "Marginal effect")
> for(i in 1:3) lines(x, dqhat[ ,i], lty = c(1, 2, 4)[i], lwd = 2)
```

The effect of solar radiation on different quantiles of ozone levels shows a nonlinear behavior, especially at lower ranges of radiation (below 50 Ly) and on the median ozone. It might be worth testing the goodness-of-fit of the model. In the previous analysis, it was found evidence of lack of fit for the linear specification (5). In contrast, the output reported below indicates that, in general, the goodness of fit of the quantile models based on the transformation model (15) has improved since the test statistics are now smaller at all values of p. However, such improvement is not yet satisfactory for the median.

```
> GOFTest(fit.rqt, alpha = 0.05, B = 1000, seed = 416)

Goodness-of-fit test for quantile regression based on the cusum process
Quantile 0.1: Test statistic = 0.0393; p-value = 0.025
Quantile 0.5: Test statistic = 0.1465; p-value = 0.005
Quantile 0.9: Test statistic = 0.0212; p-value = 0.127
```

The TS and RC estimators generally provide similar estimates and predictions. However, computation based on the cusum process tends to be somewhat slow, as shown further below. This is also true for the RC test provided by `GOFTest()`.

```
> system.time(fit.rqt <- rcrq(Ozone ~ Solar.R, data = dd, tsf = "mcjI",
+    symm = TRUE, dbounded = FALSE, lambda = seq(1, 3, by = 0.005),
+    tau = c(.1, .5, .9)))
```

```
   user  system elapsed
  36.88    0.03   37.64
```

An example using doubly bounded transformations is demonstrated using the A-level Chemistry Scores data set. The latter is available from **Qtools** and it consists of 31022 observations of A-level scores in Chemistry for England and Wales students, 1997. The data set also includes information of prior academic achievement as assessed with General Certificate of Secondary Education (GCSE) average scores. The goal is to evaluate the ability of GCSE to predict A-level scores. The latter are based on national exams in specific subjects (e.g. chemistry) with grades ranging from A to F. For practical purposes, scores are converted numerically as follows: A = 10, B = 8, C = 6, D = 4, E = 2, and F = 0. The response is therefore doubly bounded between 0 ad 10. It should be noted that this variable is discrete, although, for the sake of simplicity, here it is assumed that the underlying process is continuous.

The model considered here is

$$Q_{h_{AOa}\{\text{score}\}}(p) = \beta_0(p) + \beta_1(p) \cdot \text{gcse}, \tag{16}$$

where h_{AOa} denotes the asymmetric Aranda-Ordaz transformation in (12). This model is fitted for $p = 0.9$:

```
> data(Chemistry)
> fit.rqt <- tsrq(score ~ gcse, data = Chemistry, tsf = "ao", symm = FALSE,
+    lambda = seq(0, 2, by = 0.01), tau = 0.9)
```

The predicted 90th centile of A-level scores conditional on GCSE and the marginal effect of GCSE are plotted in Figure 6. There is clearly a positive, nonlinear association between the two scores. The nonlinearity is partly explained by the floor and ceiling effects which result from the boundedness of the measurement scale. Note, however, that the S-shaped curve is not symmetric about the inflection point. As a consequence, the marginal effect is skewed to the left. Indeed, the estimate $\hat{\lambda}_{0.9} = 0$ and the narrow confidence interval give support to a complementary log-log transformation:

```
> summary(fit.rqt, conditional = FALSE, se = "nid")

call:
summary.rqt(object = fit.rqt, se = "nid", conditional = FALSE)

Aranda-Ordaz asymmetric transformation (doubly bounded response)

Summary for unconditional inference

tau =  0.9

Optimal transformation parameter:
       Value   Std. Error  Lower bound  Upper bound
```

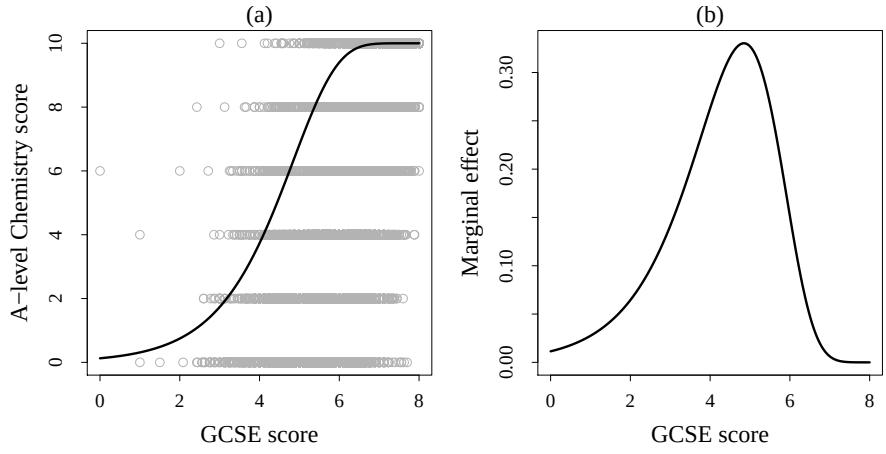

Figure 6: Predicted 90th centile of A-level scores conditional on GCSE scores (a) and corresponding estimated marginal effect (b) using the asymmetric Aranda-Ordaz transformation in the A-level Chemistry Scores data set.

```
 0.000000000  0.001364422 -0.002674218  0.002674218

Coefficients linear model (transformed scale):
                Value  Std. Error Lower bound Upper bound
(Intercept) -4.3520060 0.015414540  -4.3822179  -4.3217941
gcse         0.8978072 0.002917142   0.8920898   0.9035247

Degrees of freedom: 31022 total; 31020 residual
```

Alternatively, one can estimate the parameter δ_p using a two-parameter transformation:

```
> coef(tsrq2(score ~ gcse, data = chemsub, dbounded = TRUE,
+    lambda = seq(0, 2, by = 0.1), delta = seq(0, 2, by = 0.1),
+    tau = 0.9), all = TRUE)

(Intercept)        gcse      lambda       delta
 -4.1442274   0.8681246   0.0000000   0.0000000
```

These results confirm the asymmetric nature of the relationship since $\hat{\delta}_{0.9} = 0$. Similar results (not shown) were obtained with `nlrq2()`.

In conclusion, the package **Qtools** offers several options in terms of transformations and estimation algorithms, the advantages and disadvantages of which are discussed by Geraci and Jones (2015). In particular, they found that the symmetric Proposal I transformation improves considerably on the Box-Cox method and marginally on the Aranda-Ordaz transformation in terms of mean squared error of the predictions. Also, asymmetric transformations do not seem to improve sufficiently often on symmetric transformations to be especially recommendable. However, the Box-Cox and the symmetric Aranda-Ordaz transformations should not be used when individual out-of-range predictions represent a potential inconvenience as, for example, in multiple imputation (see section further below). Finally, in some situations transformation-based quantile regression may be competitive as compared to methods based on smoothing, as demonstrated by a recent application to anthropometric charts (Boghossian et al., 2016).

Conditional LSS

Quantile-based measures of location, scale, and shape can be assessed *conditionally* on covariates. A simple approach is to a fit a linear model as in (4) or a transformation-based model as in (7), and then predict $\hat{Q}_{Y|X}(p)$ to obtain the conditional LSS measures in Equation 3 for specific values of $\mathbf{x}$.

Estimation of conditional LSS can be carried out by using the function `qlss.formula()`. The conditional model is specified in the argument `formula`, while the probability p is given in `probs`. (As seen in Equation 3, the other probabilities of interest to obtain the decomposition of the conditional quantiles are $1 - p$, 0.25, 0.5, and 0.75.) The argument `type` specifies the required type of regression model, more specifically `"rq"` for linear models and `"rqt"` for transformation-based models. The function `qlss.formula()` will take any additional argument to be passed to `quantreg::rq()` or `tsrq()` (e.g. `subset`, `weights`, etc.).

Let's consider the New York Air Quality data example discussed in the previous section and assume that the transformation model (15) holds for the quantiles $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. Then the conditional LSS summary of the distribution of ozone conditional on solar radiation for $p = 0.05$ and $p = 0.1$ is calculated as follows:

```
> fit.qlss <- qlss(formula = Ozone ~ Solar.R, data = airquality, type =
+    "rqt", tsf = "mcjI", symm = TRUE, dbounded = FALSE,  lambda =
+    seq(1, 3, by = 0.005), probs = c(0.05, 0.1))
> fit.qlss

call:
qlss.formula(formula = Ozone ~ Solar.R, probs = c(0.05, 0.1),
    data = airquality, type = "rqt", tsf = "mcjI", symm = TRUE,
    dbounded = FALSE, lambda = seq(1, 3, by = 0.005))

Conditional Quantile-Based Location, Scale, and Shape
-- Values are averaged over observations --

** Location **
Median
```

```
[1] 30.2258
** Scale **
Inter-quartile range (IQR)
[1] 43.40648
Inter-quantile range (IPR)
    0.05      0.1
88.02909 73.93430
**Shape**
Skewness index
     0.05       0.1
0.5497365 0.5180108
Shape index
    0.05      0.1
1.960315 1.661648
```

The output, which is of class `"qlss"`, is a named list with the same LSS measures seen in the case of unconditional quantiles. However, these are now conditional on solar radiation. By default, the predictions are the fitted values, which are averaged over observations for printing purposes. An optional data frame for predictions can be given via the argument `newdata` in `predict.qlss()`. If `interval = TRUE`, the latter computes confidence intervals at the specified `level` using `R` bootstrap replications (it is, therefore, advisable to set the seed before calling `predict.qlss()`). The conditional LSS measures can be conveniently plotted using the `plot.qlss()` function as shown in the code below. The argument `z` is required and specifies the covariate used for plotting. Finally, the argument `whichp` specifies one probability (and one only) among those given in `probs` that should be used for plotting (e.g. $p = 0.1$ in the following example).

```
> set.seed(567)
> x <- seq(9, 334, length = 200)
> qhat <- predict(fit.qlss, newdata = data.frame(Solar.R = x),
+    interval = TRUE, level = 0.90, R = 500)
> plot(qhat, z = x, whichp = 0.1, interval = TRUE, type = "l",
+    xlab = "Solar radiation (lang)", lwd = 2)
```

Figure 7 shows that both the median and the IQR of ozone increase nonlinearly with increasing solar radiation. The distribution of ozone is skewed to the right and the degree of asymmetry is highest at low values of solar radiation. This is due to the extreme curvature of the median which takes on values close to the 10th centile (Figure 5). (Recall that the index approaches 1 when $Q(p) \approx Q(0.5)$.) However, the sparsity of observations at the lower end of the observed range of solar radiation determines substantial uncertainty as reflected by the wider confidence interval (Figure 7). At $p = 0.1$, the conditional shape index is on average equal to 1.66 and it increases monotonically from 1.32 to about 1.85, remaining always below the tail-weight threshold of a normal distribution (1.90).

Other functions in Qtools

Restricted quantile regression

Besides a loss of precision, high sparsity (low density) might also lead to a violation of the basic property of monotonicity of quantile functions. Quantile crossing occurs when $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(p) > \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(p')$ for some $\mathbf{x}_i$ and $p < p'$. This problem typically occurs in the outlying regions of the design space (Koenker, 2005) where also sparsity occurs more frequently. Balanced designs with larger sample sizes would then offer some assurance against quantile crossing, provided, of course, that the QR models are correctly specified. Model misspecification, indeed, can still be a cause of crossing of the quantile curves. Restricted regression quantiles (RRQ) (He, 1997) might offer a practical solution when little can be done in terms of modelling. This approach applies to a subclass of linear models

$$Y = \mathbf{x}^\top \beta + \epsilon$$

and linear heteroscedastic models

$$Y = \mathbf{x}^\top \beta + \left(\mathbf{x}^\top \gamma\right) \epsilon,$$

where $\mathbf{x}^\top \gamma > 0$ and $\epsilon \sim F$. Basically, it consists in fitting a reduced regression model passing through the origin. The reader is referred to He (1997) for details. Here, it is worth stressing that when the restriction does not hold, i.e. if the model is more complex than a location–scale-shift model, then RRQ may yield unsatisfactory results He (1997). See also Zhao (2000) for an examination of the asymptotic

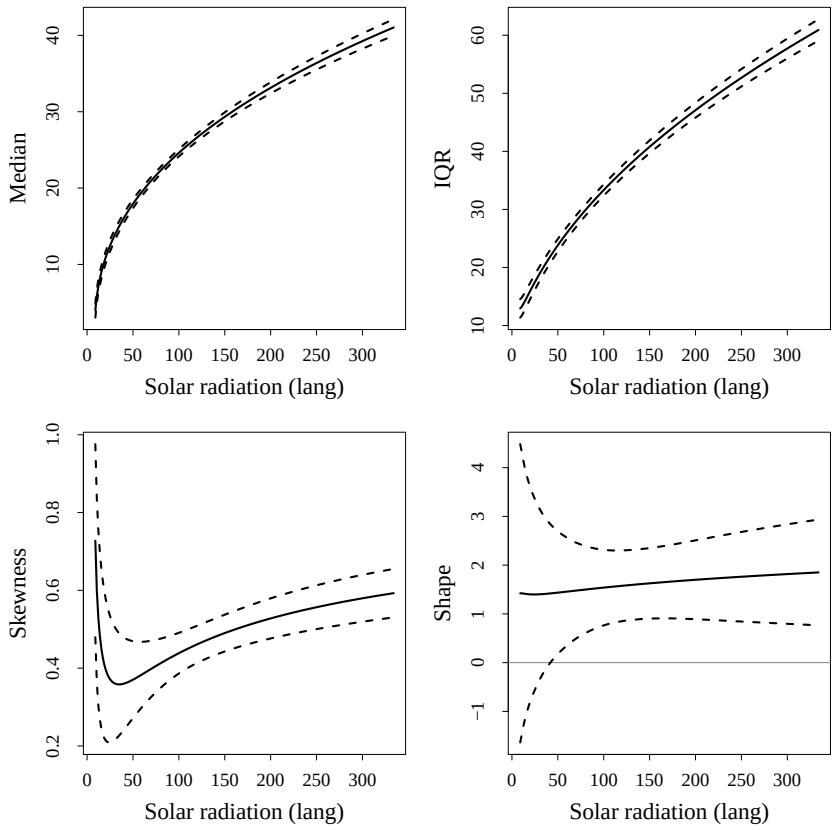

Figure 7: Location, scale and shape of ozone levels conditional on solar radiation in the New York Air Quality data set. Dashed lines denote the bootstrapped 90% point-wise confidence intervals.

properties of the restricted QR estimator. In particular, the relative efficiency of RRQ as compared to RQ depends on the error distribution. For some common unimodal distributions, Zhao (2000) showed that RRQ in iid models is more efficient than RQ. This property is lost when the error is asymmetric. In contrast, the efficiency of RRQ in heteroscedastic models is comparable to that of RQ even for small samples.

The package **Qtools** provides the functions `rrq()`, `rrq.fit()` and `rrq.wfit()` which are, respectively, the *restricted* analogs of `rq()`, `rq.fit()`, and `rq.wfitv` in **quantreg**. S3 methods `print()`, `coef()`, `predict()`, `fitted()`, `residuals()`, and `summary()` are available for objects of class `"rrq"`. In particular, confidence intervals are obtained using the functions `boot()` and `boot.ci()` from package **boot**. Future versions of the package will develop the function `summary.rrq()` to include asymptotic standard errors (Zhao, 2000). An application is shown below using an example discussed by Zhao (2000). The data set, available from **Qtools**, consists of 118 measurements of esterase concentrations and number of bindings counted in binding experiments.

```
> data("esterase")
> taus <- c(.1, .25, .5, .75, .9)
> fit.rq <- rq(Count ~ Esterase, data = esterase, tau = taus)
> yhat1 <- fitted(fit.rq)
> fit.rrq <- rrq(Count ~ Esterase, data = esterase, tau = taus)
> yhat2 <- fitted(fit.rrq)
```

The predicted 90th centile curve crosses the 50th and 75th curves at lower esterase concentrations (Figure 8). The crossing is removed in predictions based on RRQs.

As discussed above, the reliability of the results depends on the validity of the restriction carried by RRQ. A quick check can be performed using the location–scale-shift specification of the Khmaladze test.

```
> kt <- KhmaladzeTest(formula = Count ~ Esterase, data = esterase,
+ taus = seq(.05,.95,by = .01), nullH = "location-scale")
> KhmaladzeFormat(kt, 0.05)
```

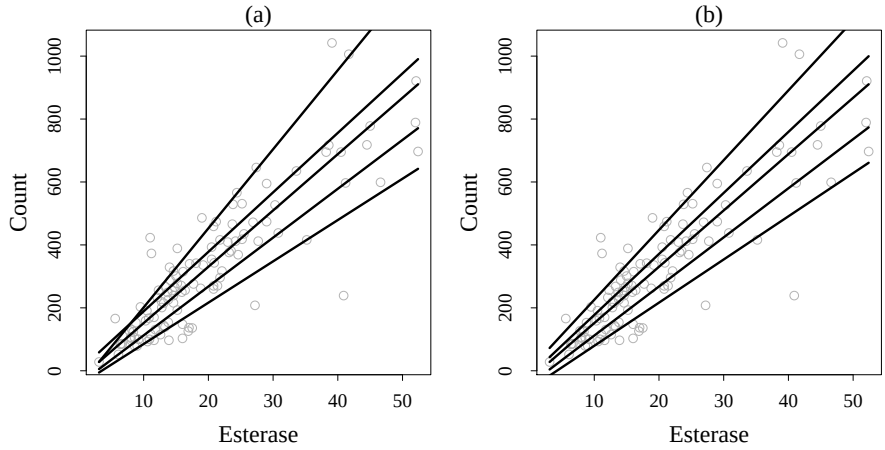

Figure 8: Predicted quantiles of number of bindings conditional on esterase concentration using regression quantiles (a) and restricted regression quantiles (b) in the Esterase data set.

```
Khmaladze test for the location-shift hypothesis
Joint test is not significant at 10% level
Test(s) for individual slopes:
 not significant at 10% level
```

The quantile crossing problem can be approached also by directly rearranging the fitted values $\hat{Q}_{Y|X=\mathbf{x}}(p)$ to obtain monotone (in p) predictions for each $\mathbf{x}$ (Chernozhukov et al., 2010). This method is implemented in the package **Rearrangement** (Graybill et al., 2016). As compared to RRQ, this approach is more general as it is not confined to, for example, location–scale-shift models (Chernozhukov et al., 2010); however, in contrast to RRQ, it does not yield estimates of parameters (e.g. slopes) of the model underlying the final monotonised curves. Such estimates, available from `"rrq"` objects, may be of practical utility when summarising the results.

Conditional quantiles of discrete data

Modelling conditional functions of discrete data is less common and, on a superficial level, might even appear as an unnecessary complication. However, a deeper look at its rationale will reveal that a distribution-free analysis can provide insightful information in the discrete case as it does in the continuous case. Indeed, methods for conditional quantiles of continuous distributions can be—and have been—adapted to discrete responses.

The package **Qtools** offers some limited functionalities for count and binary data. Further research is needed to develop the theory of QR for discrete data and to improve computational algorithms. Therefore, the user should use these functions with caution.

Let Y be a count variable such as, for example, the number of car accidents during a week or the number of times a patient visits their doctor during a year. As usual, X denotes a vector of covariates. Poisson regression, which belongs to the family of generalised linear models (GLMs), is a common choice for this kind of data, partly because of its availability in many statistical packages. Symbolically, $Y \sim \text{Pois}(\theta)$, where $\theta \equiv \mathbb{E}(Y|X=\mathbf{x}) = h^{-1}\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$ and h is the logarithmic link function. Note that the variance also is equal to θ. Indeed, moments of order higher than 2 governing the shape of the distribution depend on the same parameter. Every component of the conditional LSS in a Poisson model is therefore controlled by θ. If needed, more flexibility can be achieved using a distribution-free approach.

Machado and Santos Silva (2005) proposed the model

$$Q_{h(Z;p)}(p) = \mathbf{x}^\top \boldsymbol{\beta}(p), \tag{17}$$

where $Z = Y + U$ is obtained by jittering Y with a $[0,1)$-uniform noise U, independent of Y and X. In principle, any monotone transformation h can be considered. Given the continuity between counts induced by jittering, standard inference for linear quantile functions (Koenker and Bassett, 1978) can be applied to fit (17). In practice, a sample of M jittered responses Z is taken to estimate $\hat{\boldsymbol{\beta}}_m(p)$, $m = 1, \ldots, M$; the noise is then averaged out, $\hat{\boldsymbol{\beta}}(p) = \frac{1}{M}\sum_m \hat{\boldsymbol{\beta}}_m(p)$.

Machado and Santos Silva's (2005) methods, including large-n approximations for standard errors, are implemented in the function `rq.counts()`. The `formula` argument specifies a linear model as

in (17), while the argument `tsf` provides the desired transformation h. By default, this is the log transformation (i.e. Box-Cox with parameter $\lambda_p = 0$) but other transformations are allowed. Note that `GOFTest()` can be applied to `"rq.counts"` objects as well.

Qtools provides functions for modelling binary responses as well. First of all, it is useful to note that the classical GLM for a binary response $Y \sim \text{Bin}(1, \pi)$ establishes a relationship between the probability $\Pr(Y = 1) = \pi$ and a set of predictors $\mathbf{x}$. The application of QR to binary outcomes relies on the continuous latent variable regression formulation

$$Y^* = \mathbf{x}^\top \boldsymbol{\beta} + \epsilon \tag{18}$$

and assumes that the binary observations are the result of the dichotomization $Y = I(Y^* > 0)$, with Y^* unobserved.

Maximum score estimation, originally developed by Manski (1975, 1985), is equivalent to estimating the conditional quantiles of the latent variable Y^*. However, the optimization problem offers numerical challenges due to the the piecewise linearity of the indicator function and the nonconvexity of the loss function. The function `rq.bin()` is the main function to obtain binary regression quantiles. It is a wrapper for the function `rqbin.fit()` which calls Fortran code written for simulated annealing estimation (Goffe et al., 1994). **Qtools** offers a limited number of functions for objects of class `"rq.bin"` including `coef()` and `predict()`. These methods should be considered still experimental. In particular, the user should be aware that the estimates obtained from the fitting procedure may be sensitive to different settings of the simulated annealing algorithm. The latter can be controlled using `rqbinControl()`.

Quantile-based multiple imputation

Regression models play an important role in conditional imputation of missing values. QR can be used as an effective approach for multiple imputation (MI) when location-shift models are inadequate (Muñoz and Rueda, 2009; Bottai and Zhen, 2013; Geraci, 2016).

In **Qtools**, `mice.impute.rq()` and `mice.impute.rrq()` are auxiliary functions written to be used along with the functions of the R package **mice** (van Buuren and Groothuis-Oudshoorn, 2011). The former is based on the standard QR estimator (`rq.fit()`) while the latter on the restricted counterpart (`rrq.fit()`). Both imputation functions allow for the specification of the transformation-based QR models discussed previously. The equivariance property is useful to achieve linearity of the conditional model and to ensure that imputations lie within some interval when imputed variables are bounded. An example is available from the help file `?mice.impute.rq` using the `nhanes` data set. See also Geraci (2016) for a thorough description of these methods.

Final remarks

Quantiles have long occupied an important place in statistics. The package **Qtools** builds on recent methodological and computational developments of quantile functions and related methods to promote their application in statistical data modelling.

Acknowledgements

This work was partially supported by an ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina. I wish to thank anonymous reviewers for their helpful comments and Alexander McLain for his help with revising the final draft of the manuscript.

Bibliography

F. J. Aranda-Ordaz. On two families of transformations to additivity for binary response data. *Biometrika*, 68(2):357–363, 1981. [p125]

A. Azzalini and A. W. Bowman. A look at some data on the Old Faithful Geyser. *Journal of the Royal Statistical Society C*, 39(3):357–365, 1990. [p122]

Y. Benjamini and A. M. Krieger. Concepts and measures for skewness with data-analytic implications. *Canadian Journal of Statistics*, 24(1):131–140, 1996. [p120]

D. F. Benoit, R. Al-Hamzawi, K. Yu, and D. V. den Poel. *bayesQR: Bayesian Quantile Regression*, 2014. URL http://CRAN.R-project.org/package=bayesQR. R package version 2.2. [p117]

N. S. Boghossian, M. Geraci, E. M. Edwards, K. A. Morrow, and J. D. Horbar. Anthropometric charts for infants born between 22 and 29 weeks' gestation. *Pediatrics*, 2016. doi:10.1542/peds.2016-1641. [p131]

M. Bottai and H. Zhen. Multiple imputation based on conditional quantile estimation. *Epidemiology, Biostatistics, and Public Health*, 10(1):e8758, 2013. [p135]

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society B*, 26 (2):211–252, 1964. [p125]

M. Buchinsky. Quantile regression, Box-Cox transformation model, and the US wage structure, 1963-1987. *Journal of Econometrics*, 65(1):109–154, 1995. [p125, 127]

A. Canty and B. D. Ripley. *boot: Bootstrap R (S-PLUS) Functions*, 2014. URL http://CRAN.R-project.org/package=boot. R package version 1.3-15. [p127]

G. Chamberlain. *Quantile Regression, Censoring, and the Structure of Wages*, volume 1. Cambridge University Press, Cambridge, UK, 1994. [p125, 127]

V. Chernozhukov, I. Fernandez-Val, and A. Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010. [p134]

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997. [p127]

H.-M. Dehbi, M. Cortina-Borja, and M. Geraci. Aranda-Ordaz quantile regression for student performance assessment. *Journal of Applied Statistics*, 43(1):58–71, 2016. [p125]

K. Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2(2):267–277, 1974. [p122]

M. Geraci. Linear quantile mixed models: The lqmm package for Laplace quantile regression. *Journal of Statistical Software*, 57(13):1–29, 2014. [p117]

M. Geraci. Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Statistical Methods in Medical Research*, 25(4):1393–1421, 2016. [p135]

M. Geraci and M. Bottai. Linear quantile mixed models. *Statistics and Computing*, 24(3):461–479, 2014. [p117]

M. Geraci and M. C. Jones. Improved transformation-based quantile regression. *Canadian Journal of Statistics*, 43(1):118–132, 2015. [p125, 126, 127, 131]

W. Gilchrist. *Statistical Modelling with Quantile Functions*. Chapman & Hall/CRC, Boca Raton, FL, 2000. [p118, 120]

W. L. Goffe, G. D. Ferrier, and J. Rogers. Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, 60(1):65–99, 1994. [p135]

W. Graybill, M. Chen, V. Chernozhukov, I. Fernandez-Val, and A. Galichon. *Rearrangement: Monotonize Point and Interval Functional Estimates by Rearrangement*, 2016. URL http://CRAN.R-project.org/package=Rearrangement. R package version 2.1. [p134]

R. A. Groeneveld. A class of quantile measures for kurtosis. *The American Statistician*, 52(4):pp. 325–329, 1998. [p119, 120]

R. A. Groeneveld and G. Meeden. Measuring skewness and kurtosis. *Journal of the Royal Statistical Society D*, 33(4):pp. 391–399, 1984. [p119]

A. Hald. *A History of Probability and Statistics and their Applications before 1750*. John Wiley & Sons, New York, NY, 2003. [p117]

X. He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997. [p132]

X. M. He and L. X. Zhu. A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, 98(464):1013–1022, 2003. [p123]

P. S. Horn. A measure for peakedness. *The American Statistician*, 37(1):55–56, 1983. [p119]

R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4): 361–365, 1996. [p118]

M. C. Jones. Connecting distributions with power tails on the real line, the half line and the interval. *International Statistical Review*, 75(1):58–69, 2007. [p125]

M. C. Jones, J. F. Rosco, and A. Pewsey. Skewness-invariant measures of kurtosis. *The American Statistician*, 65(2):89–95, 2011. [p119, 120]

E. V. Khmaladze and H. L. Koul. Martingale transforms goodness-of-fit tests in regression models. pages 995–1034, 2004. [p123]

R. Koenker. *Quantile Regression*. Cambridge University Press, New York, NY, 2005. [p117, 123, 127, 132]

R. Koenker. *quantreg: Quantile Regression*, 2013. URL http://CRAN.R-project.org/package=quantreg. R package version 5.05. [p117]

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. [p122, 123, 134]

R. Koenker and V. D'Orey. Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society C*, 36(3):383–393, 1987. [p117]

R. Koenker and J. A. F. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999. [p123]

R. Koenker and B. J. Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1-2):265–283, 1996. [p117, 125]

R. Koenker and Z. J. Xiao. Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612, 2002. [p122, 123]

E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, CA, 1975. [p122]

Y. Ma, M. G. Genton, and E. Parzen. Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, 63(2):227–243, 2011. [p117, 118]

J. A. F. Machado and J. Mata. Box–Cox quantile regression and the distribution of firm sizes. *Journal of Applied Econometrics*, 15(3):253–274, 2000. [p127]

J. A. F. Machado and J. M. C. Santos Silva. Quantiles for counts. *Journal of the American Statistical Association*, 100(472):1226–1237, 2005. [p117, 134]

C. F. Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228, 1975. [p135]

C. F. Manski. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333, 1985. [p135]

Y. M. Mu and X. M. He. Power transformation toward a linear regression quantile. *Journal of the American Statistical Association*, 102(477):269–279, 2007. [p125, 127]

J. F. Muñoz and M. Rueda. New imputation methods for missing data using quantiles. *Journal of Computational and Applied Mathematics*, 232(2):305–317, 2009. [p135]

E. Parzen. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74 (365):105–121, 1979. [p118]

E. Parzen. Quantile probability and statistical data modeling. *Statistical Science*, 19(4):652–662, 2004. [p117, 118]

J. L. Powell. *Estimation of Monotonic Regression Models Under Quantile Restrictions*, pages 357–384. Cambridge University Press, New York, NY, 1991. [p125, 127]

J. L. Powell. *Estimation of Semiparametric Models*, volume Volume 4, chapter 41, pages 2443–2521. Elsevier, 1994. [p123]

D. Ruppert. What is kurtosis?: An influence function approach. *The American Statistician*, 41(1):1–5, 1987. [p119]

R. M. Sakia. The Box–Cox transformation technique: A review. *Journal of the Royal Statistical Society D*, 41(2):169–178, 1992. [p125]

L. Smith and B. Reich. *BSquare: Bayesian Simultaneous Quantile Regression*, 2013. URL http://CRAN.R-project.org/package=BSquare. R package version 1.1. [p117]

R. G. Staudte. Inference for quantile measures of kurtosis, peakedness and tail-weight. *arXiv preprint arXiv:1047.6461v1 [math.ST]*, 2014. doi: 10.1080/03610926.2015.1056366. [p119]

J. W. Tukey. Which part of the sample contains the information? *Proceedings of the National Academy of Sciences of the United States of America*, 53(1):127–134, 1965. [p118]

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. [p135]

K. M. Yu and M. C. Jones. Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237, 1998. [p125]

Q. S. Zhao. Restricted regression quantiles. *Journal of Multivariate Analysis*, 72(1):78–99, 2000. [p132, 133]

J. X. Zheng. A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, 14(1):123–138, 1998. [p123]

Marco Geraci
Department of Epidemiology and Biostatistics
Arnold School of Public Health, University of South Carolina
915 Greene Street, Columbia SC 29204
United States of America geraci@mailbox.sc.edu