Qtools: A Useful Package for Quantiles

Marco Geraci
University of South Carolina

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.

Keywords: discrete random variables, goodness of fit, imputation, location-scale-shape measures, transformations.

1. Applications and theory of quantiles

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 and Bassett 1985; 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. However, it was Sir Francis Galton with his remarkable studies on anthropometry to provide a more systematic definition of quantiles and to popularize terms such as 'median', 'quartile' and 'percentile' (Galton 1882, 1885). He was also responsible for introducing the ogive (Galton 1875), which today it is called 'quantile function'. (However, it is not uncommon to find that the ogive is used as synonym with cumulative distribution function.) Over time, the interest in the applications of quantiles has grown in parallel with advances in the corresponding inference and computing algorithms.

The theoretical studies on unconditional and conditional quantiles of continuous random variables started to appear in the statistical literature of the 20th century. According to David (1995), it seems that the term 'quantile' itself was first used in print by Kendall (1940). In his paper, Kendall showed that sample quantiles of independent and identically distributed (*iid*) observations have an asymptotic normal distribution (see also Hald, 1998, for historical notes on asymptotic studies on unconditional quantiles). The asymptotic theory of conditional quantile functions is more recent, beginning from the work of Bassett and Koenker (1978) on the median estimator and its generalization to other quantiles by Koenker and Bassett (1978).

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, Genton, and Parzen (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.

Finally, it is worth mentioning that asymptotic results are also available for unconditional (see for example Oberhofer and Haupt 2005, and references therein) and conditional (Parente and Santos Silva 2015) sample quantiles when data are not independent, as well as for nonlinear median functions (Wang 1995).

The focus of this paper is on the package **Qtools**, a collection of utilities for unconditional and conditional quantiles, written for the R statistical computing environment (R Core Team 2015) and available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=Qtools`.

2. Unconditional quantiles

2.1. Definition of quantiles and their properties

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 $Q(p) = \inf_y\{F_Y(y) \geq p\}$, $0 < p < 1$. (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 one-to-one relationship between the values of Y and the probability p is lost where the distribution function is piecewise constant and, by convention, the quantile p is defined 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). The advantages of using quantiles rather than probabilities lie in their 'algebraic' properties which can be summarized in:

1. $Q_{a+bY}(p) = a + bQ_Y(p)$, $a \in \mathbb{R}$, $b \in \mathbb{R}_+$ (**location-shift rule**);
2. if $Q_1(p)$ and $Q_2(p)$ are two QFs, then $Q_1(p) + Q_2(p)$ is a QF (**addition rule**);
3. if $Q_1(p)$ and $Q_2(p)$ are two QFs, then $\pi Q_1(p) + (1-\pi)Q_2(p)$, $0 \leq \pi \leq 1$, lies between the two distributions (**intermediate rule**);
4. if $Q_1(p)$ and $Q_2(p)$ are two positive QFs, then $Q_1(p) \cdot Q_2(p)$ is a QF (**multiplication rule**);
5. if $Q_Y(p)$ is the QF of Y, the reflection $-Q_Y(1-p)$ is the QF of $-Y$ (**reflection rule**);
6. if $Q_Y(p)$ is the QF of Y, the reciprocal $1/Q_Y(1-p)$ is the QF of $1/Y$ (**reciprocal rule**);
7. if $h(\cdot)$ is a non-decreasing function on $\mathbb{R}$, then $Q_{h(Y)}(p) = h\{Q_Y(p)\}$. Hence $Q_Y(p) = h^{-1}\{Q_{h(Y)}(p)\}$ (**Q-tranformation rule** or equivariance to monotone transformations).

(Note that, in general, the last property does not hold for the expected value, i.e. $\mathsf{E}\{h(Y)\} \neq h\{\mathsf{E}(Y)\}$.)

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** (R Core Team 2015) 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). Alternatively, one could consider a model for F (or Q), indexed by some low-dimensional parameter, say θ. Estimation of θ can be carried out efficiently using one of several available methods, such as maximum likelihood estimation (MLE), the method of moments and the method of percentiles (Gilchrist 2000). For instance, suppose that Y follows an exponential distribution

$$F_Y(y;\theta) = 1 - e^{-\theta y},$$

with rate $\theta \in \mathbb{R}_+$. The corresponding quantile function is given by

$$Q_Y(p;\theta) = -\frac{\log(1-p)}{\theta}.$$

Once the parameter's estimate $\hat{\theta}$ is obtained, it is straightforward to predict $\hat{Q}_Y(p;\hat{\theta})$ for any $0 < p < 1$.

Distribution-free and distribution-based quantile estimation obviously differ in some respects. A major distinction can be made in terms of the possible range of predictions. In contrast to a distribution-free estimator, a distribution-based approach will yield predictions that are not constrained to lie between the first and last order sample statistics. However, extrapolation outside the observed range of values should always be taken with a pinch of salt. The main focus of the package **Qtools** is on the distribution-free approach, though some on-going developments of distribution-based quantile methods will be described in the next sections.

2.2. Sample quantiles and large-n properties

Let $\mathcal{Y}_n = (Y_1, Y_2, \ldots, Y_n)$ be an *iid* sample of size n from the population F_Y. Let ξ_p denote the pth population quantile and $\hat{\xi}_p$ the corresponding sample quantile based on the sample $\mathcal{Y}_n$. (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 are independent with common ξ_p but different distribution function, i.e. $Y_i \sim F_{Y_i}$. In this case the variance depends on the density f_{Y_i}, i.e.

$$\omega_i^2 = \frac{p(1-p)}{\{f_{Y_i}(\xi_p)\}^2}. \tag{2}$$

The density evaluated at the pth quantile, $f(\xi_p)$, is called density-quantile function by Parzen (1979). Its reciprocal, $s(p) \equiv 1/f(\xi_p)$, is called sparsity function (Tukey 1965) or quantile-density function (Parzen 1979). It is easy to verify that $s(p) = dQ(p)/dp$.

As mentioned in the previous section, 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 $F_Y^{mid}(y) = F_Y(p) - 0.5p_Y(y)$, where $p_Y(y)$ denotes the probability mass function (Parzen 2004). In particular, they showed that, as n becomes large, $\sqrt{n}\left(\hat{\xi}_p - \xi_p\right)$ is approximately normal with mean 0. Under *iid* assumptions, the expression for the variance ω^2 is similar to that in (1); see Ma *et al.* (2011) for details.

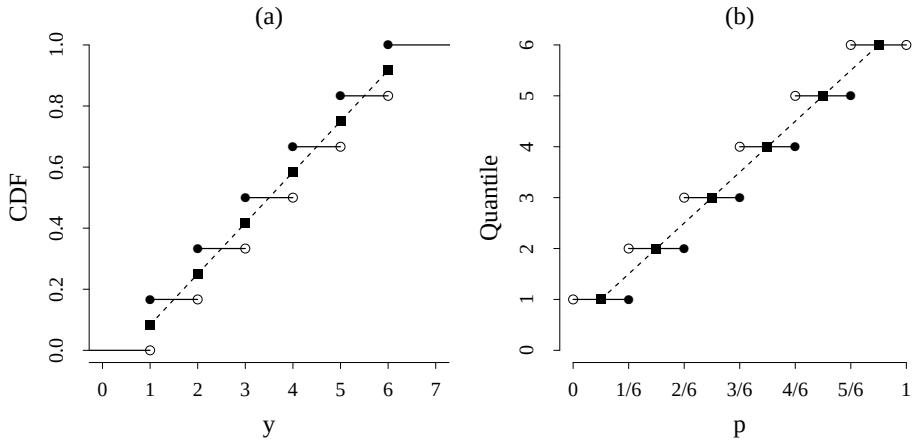

Figure 1: Cumulative distribution (a) and quantile (b) functions for the 'die rolling' experiment. The ordinary CDF and quantile function are represented by step-functions (solid lines), with the convention that, at the point of discontinuity or 'jump', the function takes its value corresponding to the ordinate of the filled circle as opposed to that of the hollow circle. The mid-CDF and mid-quantile functions are represented by piecewise linear functions (dashed lines) connecting the filled squares.

An illustration based on the outcome of rolling a fair die is given in Figure 1. In this experiment, the variable Y is discrete and each of the six values has a probability 1/6. The mid-distribution and mid-quantile functions interpolate between the steps of, respectively, the ordinary CDF and quantile functions.

The package **Qtools** provides the functions `midecdf` and `midquantile`, which return an object of class '`list`' containing x and y coordinates, along with the corresponding interpolating function as an attribute named '`function`'. This is shown in the example below.

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

$x
 [1]  0  1  2  3  4  5  6  7  8  9 10 12
```

```
$y
 [1] 0.0110 0.0530 0.1555 0.3305 0.5365 0.7175 0.8450 0.9215 0.9635 0.9860
 [2] 0.9965 0.9995
attr(,"function")

[...]

R> xmid

$x
 [1] 0.0110 0.0530 0.1555 0.3305 0.5365 0.7175 0.8450 0.9215 0.9635 0.9860
 [2] 0.9965 0.9995
$y
 [1]  0  1  2  3  4  5  6  7  8  9 10 12

attr(,"function")

[...]
```

A confidence interval for sample mid-quantiles can be obtained using `midquantile.ci`. This function returns an object of class '`data.frame`' containing sample mid-quantiles, lower and upper bounds of the confidence interval 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.

```
R> x <- midquantile.ci(y, probs = 1:3/4, level = 0.95)
R> x

    midquantile    lower    upper
25%    2.540000 2.416462 2.663538
50%    3.822816 3.693724 3.951907
75%    5.254902 5.072858 5.436946

R> attr(x, "stderr")

[1] 0.06295447 0.06578432 0.09276875
```

2.3. 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, Rosco, and Pewsey 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:

$$\underbrace{\frac{IPR(p)}{IPR(r)}}_{\text{kurtosis}} = \underbrace{\frac{IPR(p)}{IPR(q)}}_{\text{tail-weight}} \cdot \underbrace{\frac{IPR(q)}{IPR(r)}}_{\text{peakedness}}, \tag{3}$$

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

$$Q(p) = \underbrace{Q(0.5)}_{\text{median}} + \frac{1}{2}\underbrace{IPR(0.25)}_{\text{IQR}} \cdot \underbrace{\frac{IPR(p)}{IPR(0.25)}}_{\text{shape index}} \cdot \left(\underbrace{\frac{Q(p) + Q(1-p) - 2Q(0.5)}{IPR(p)}}_{\text{skewness index}} - 1 \right). \tag{4}$$

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 (4) is loosely defined as 'shape index' (Gilchrist 2000, p.72), although it can be seen as the tail-weight measure given in (3) when $p < 0.25$. For symmetric distributions, the contribution of the skewness index (Gilchrist 2000, p.53) 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 standardized moment which can be infinite or even undefined).

The function `qlss` provides a quantile-based LSS summary as defined in (4) 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.).

```
R> qlss(fun = "qnorm", probs = 0.1) # equivalent to qlss()

$location
$location$median
```

```
[1] 0

$scale
$scale$IQR
[1] 1.34898

$scale$IPR
     0.1
2.563103

$shape
$shape$skewness
0.1
  0

$shape$shape
     0.1
1.900031

attr(,"class")
[1] "qlss"
```

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 estimated density is plotted in Figure 2, along with the mid-quantile function. The distribution is bimodal with peaks at around 54 and 80 minutes. Note that `qlss` takes the argument `type` as well as any other argument for the function `quantile`.

```
R> y <- faithful$waiting
R> qlss(y, probs = c(0.05,0.1,0.25), type = 7)

$location
$location$median
[1] 76

$scale
```

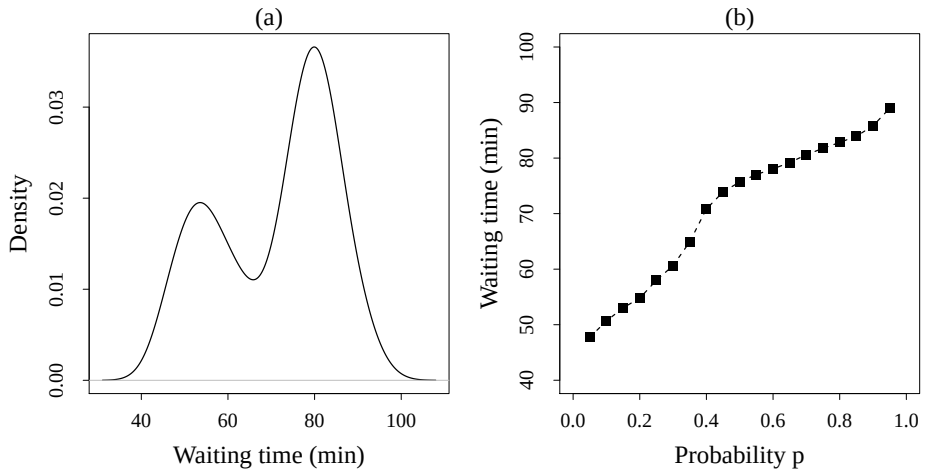

Figure 2: Estimated density (a) and empirical mid-quantile (b) functions of waiting time between eruptions in the Old Faithful Geyser data set.

```
$scale$IQR
[1] 24

$scale$IPR
0.05  0.1 0.25
  41   35   24

$shape
$shape$skewness
      0.05        0.1       0.25
-0.3658537 -0.4285714 -0.5000000

$shape$shape
    0.05      0.1     0.25
1.708333 1.458333 1.000000


attr(,"class")
[1] "qlss"
```

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.

3. Conditional quantiles

3.1. Definition of conditional quantiles

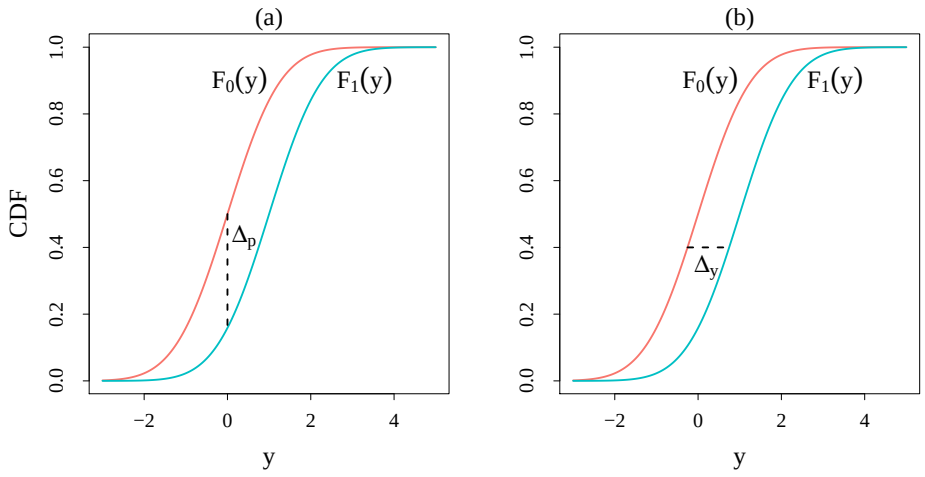

Figure 3: (a) Probability difference (à la Kolmogorov–Smirnov); (b) Quantile difference (à la Lehmann–Doksum).

Conditional modeling of quantiles can be easily illustrated starting from the simple case in which two populations are contrasted in relation to some characteristic Y. Let $F_0(y)$ and $F_1(y)$ denote the cumulative distribution functions (CDFs) of Y in the two populations, and let $\mu_0 = \int_{-\infty}^{\infty} y \, \mathrm{d}F_0(y)$ and $\mu_1 = \int_{-\infty}^{\infty} y \, \mathrm{d}F_1(y)$ be their respective means. The two CDFs are plotted in Figure 3. The question of whether and how F_1 differs from F_0 can be approached by comparing, for any fixed value y, the distance between the CDFs along the probability (vertical) axis. Therefore

$$\Delta_p(y) = F_1(y) - F_0(y).$$

A discrepancy between the two distributions at some value y would result in a positive or negative $\Delta_p(y)$. The largest (in magnitude) of all the differences $\Delta_p(y)$ is used to calculate the Kolmogorov–Smirnov test statistic. Plot (a) in Figure 3 shows the case in which F_1 is stochastically dominated by F_0, thus $\Delta_p(y) < 0$ for $-\infty < y < \infty$. A complementary approach consists in comparing, for any fixed value p, the distance of the CDFs along the quantile (horizontal) axis. Therefore

$$\Delta_y(p) = F_1^{-1}(p) - F_0^{-1}(p)$$

gives the contrast between the pth quantiles of the two distributions. This is also called 'quantile treatment effect' (Doksum 1974; Lehmann 1975; Koenker and Xiao 2002). An illustration is given in plot (b), Figure 3.

Now, define the indicator variable

$$X = \begin{cases} 1 & \text{if } Y \sim F_1, \\ 0 & \text{if } Y \sim F_0, \end{cases}$$

and the conditional distribution of Y given X

$$F_{Y|X} = \begin{cases} F_1 & \text{if } X = 1, \\ F_0 & \text{if } X = 0. \end{cases}$$

Then the mean regression model

$$\mathsf{E}(Y|X = x) = \underbrace{\mu_0}_{\text{intercept}} + \underbrace{(\mu_1 - \mu_0)}_{\text{slope}} x \tag{5}$$

can be seen arising from the contrast between F_0 and F_1. Indeed, regression analysis revolves around the modeling of differences between populations. The regression 'slope' can be written in terms of either the conditional distribution function $F_{Y|X}$ or the *conditional quantile function* $Q_{Y|X} \equiv F^{-1}_{Y|X}$, that is

$$\mu_1 - \mu_0 \equiv \int_{-\infty}^{\infty} y \left\{ \mathrm{d}F_{Y|X=1} - \mathrm{d}F_{Y|X=0} \right\} = \int_0^1 \left\{ Q_{Y|X=1}(p) - Q_{Y|X=0}(p) \right\} \mathrm{d}p. \tag{6}$$

The classical regression model for the mean thus takes an average of the quantile differences $\Delta_y(p)$'s over p. Often, the question of interest is whether the quantile differences are constant over p (i.e., the shift of the CDF is uniform along the horizontal axis) and therefore whether the mean provides an exhaustive summary of $\Delta_y(p)$, $0 < p < 1$.

The pth linear quantile regression (QR) model for Y conditional on X can be specified as

$$Q_{Y|X=x}(p) = \underbrace{Q_{Y|X=0}(p)}_{\text{intercept}} + \underbrace{\left\{ Q_{Y|X=1}(p) - Q_{Y|X=0}(p) \right\}}_{\text{slope}} x. \tag{7}$$

The intercept is now the pth quantile of $Y|X = 0$ and the slope represents the shift $\Delta_y(p)$ for a given p.

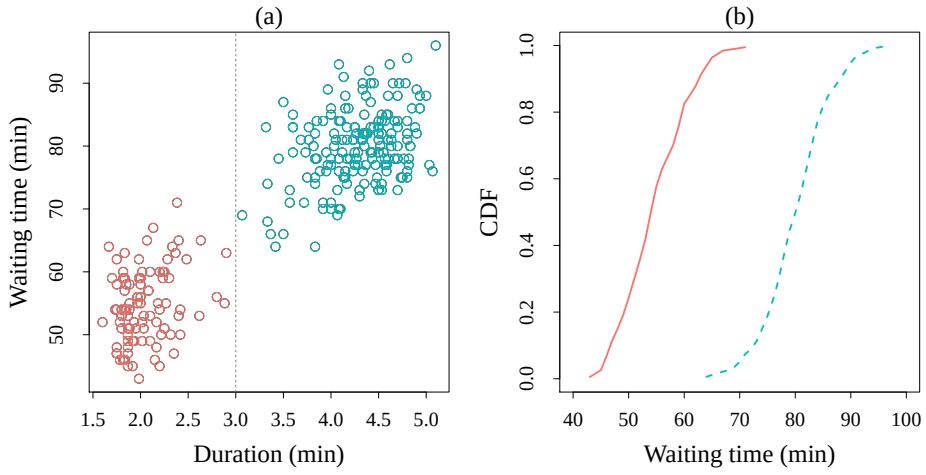

Figure 4: (a) Waiting times between eruptions against durations of eruptions (dashed vertical line drawn at 3 minutes) in the Old Faithful Geyser data set. (b) Mid-CDF of waiting time by duration of eruption (solid line, shorter than 3 minutes; dashed line, longer than 3 minutes).

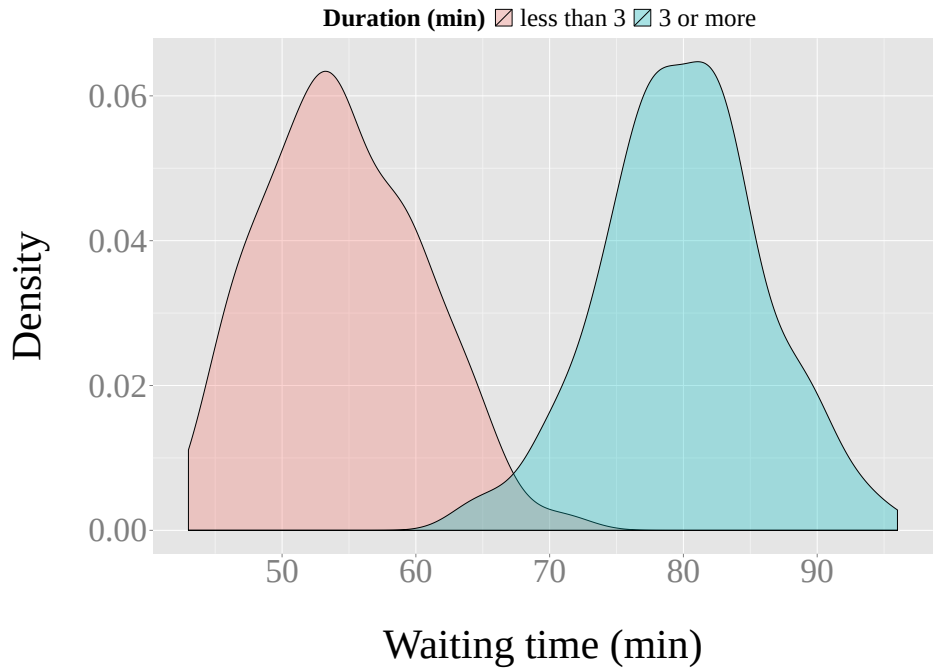

Figure 5: Estimated density of waiting time between eruptions in the Old Faithful Geyser data set, stratified by duration of eruption.

Waiting times between eruptions are plotted against the durations of the eruptions in Figure 4. 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 illustrated in Figure 5 using the estimated density function. Such bimodality is also apparent in the sample quantile function (Figure 2), which is initially convex up to about $p = 0.35$, then concave until $p = 0.65$, and then convex again. Unsurprisingly, the proportion of observations with duration less than 3 minutes is approximately 35%.

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 7. The latter is then fitted to the Old Faithful Geyser data using the function `rq` in the R package **quantreg** (Koenker 2013). The coefficient for `x`, therefore, is an estimate of the quantile difference $\Delta_y(p)$. The latter is calculated for $p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ in the following example:

```
R> require("quantreg")
R> x <- as.numeric(faithful$eruptions >= 3)
R> fit <- rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9))
R> 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 for by using the Khmaladze test as implemented in the **quantreg** package. The critical values of the test and corresponding significance levels are not readily available in the same package. These are provided by the function `KhmaladzeFormat` in the **Qtools** package.

```
R> kt <- KhmaladzeTest(formula = y ~ x, taus = seq(.05,.95,by = .01),
+ nullH = "location")
R> kt

$nullH
[1] "location"

$Tn
[1] 1.867598

$THn
[1] 1.867598

attr(,"class")
[1] "KhmaladzeTest"

R> 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
```

3.2. Inference for conditional quantiles

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

$$Q_{Y|X}(p) = \mathbf{x}^\top \boldsymbol{\beta}(p) \tag{8}$$

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

$$\frac{\partial Q_{Y|X}(p)}{\partial x_j} = \beta_j(p).$$

As in the case of unconditional quantiles, one can model the conditional quantile function using either a distribution-based or a distribution-free approach. For example, let Y follow an exponential distribution, $Y \sim \text{Exp}(\theta)$, where $\theta = \mathbf{x}^\top \boldsymbol{\gamma}$ for some k-dimensional parameter $\boldsymbol{\gamma}$. As shown in Section 2, the quantile function of Y is given by

$$Q_{Y|X}(p;\theta) = -\left\{\mathbf{x}^\top \boldsymbol{\gamma}\right\}^{-1} \log(1-p).$$

Fitting this model requires estimating $\boldsymbol{\gamma}$, e.g. using an MLE approach. Alternatively, one can define the quantile function of $1/Y$, $Y > 0$, given by (reciprocal rule)

$$\left\{Q_{Y|X}(1-p;\theta)\right\}^{-1} = -\frac{\mathbf{x}^\top \boldsymbol{\gamma}}{\log(p)}.$$

The function above, which is a valid quantile function, can be re-written in the form of (8) as $Q_{Y^{-1}|X}(\bar{p};\theta) = \mathbf{x}^\top \boldsymbol{\beta}(\bar{p})$, where $\bar{p} = 1-p$ and $\boldsymbol{\beta}(\bar{p}) \equiv \boldsymbol{\gamma}\{-\log(1-\bar{p})\}^{-1}$. (Note that the model's parameter is now a function of $\bar{p}$.) This model is now amenable to estimation based on linear programming (LP) algorithms (Koenker and Bassett 1978) 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.

Figure 6 shows data simulated from an exponential distribution $Y \sim \text{Exp}(\theta)$, where $\theta = 1 + 2x$ and x is a standard uniform. The MLE was obtained using the `optim` function. The predicted quantiles of Y conditional on x, for $p \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$, approximate the true quantile functions sufficiently well already at $n = 300$ and, as expected, the accuracy increases with increasing sample size.

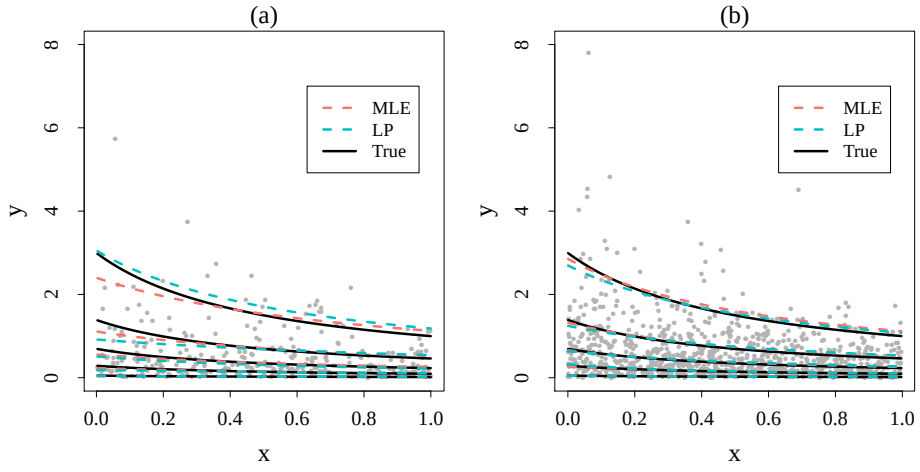

Figure 6: Quantile functions fitted by maximum likelihood estimation (MLE) and linear programming (LP) for a bivariate sample of size 300 (a) and 1000 (b).

Methods for conditional quantiles, including estimation and standard error calculation, are available in the excellent **quantreg** package, while the package **Qtools** provides some additional functionalities.

Equations 1 and 2 show that (the large-n approximation of) the sampling variance of unconditional sample quantiles is inversely proportional to the density of Y (or Y_i). Similarly, $\sqrt{n}\left(\hat{\beta}(p) - \beta(p)\right)$ has a limiting centered normal distribution whose scale depends on whether the observations are assumed to be *iid* or *nid*.

Let $f_{Y|X}$ denote the density of Y conditional on X and $\xi_p \equiv \mathbf{x}^\top \beta(p)$. In the *nid* case, the asymptotic variance-covariance matrix of p-th regression quantile is defined by the sandwich variance estimator (Koenker 2005, p.74)

$$V = p\,(1-p)\,H^{-1} D H^{-1} \tag{9}$$

where $D = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$ and $H = \lim_{n\to\infty} \frac{1}{n} \mathbf{x}_i \mathbf{x}_i^\top f_{Y_i|X_i}(\xi_p)$. In the *iid* case, the conditional density is constant and the above expression simplifies to

$$V = \omega^2 D^{-1} \tag{10}$$

where $\omega^2 = \frac{p(1-p)}{\{f_{Y_i|X_i}(\xi_p)\}^2}$, $i = 1, \ldots, n$.

If the model includes a single binary predictor x, the estimation of V can be carried out efficiently using

$$\hat{D} = \begin{bmatrix} n & n_1 \\ n_1 & n_1 \end{bmatrix}$$

and

$$\hat{H} = \begin{bmatrix} n_0 \hat{f}_0 + n_1 \hat{f}_1 & n_1 \hat{f}_1 \\ n_1 \hat{f}_1 & n_1 \hat{f}_1 \end{bmatrix},$$

where $n_1 = \sum_{i=1}^{n} x_i$ and $n_0 = n - n_1$. The density $\hat{f}_x \equiv \hat{f}_{Y|X=x}(\hat{\xi}_p)$, $x = 0, 1$, can be calculated using the reciprocal of the sparsity function. The latter is estimated using the finite difference quotient $\hat{s}(p) = d\hat{Q}(p)/2\epsilon_n$, $x = 0, 1$, where

$$d\hat{Q}(p) = \begin{cases} \hat{Q}_{Y|X=1}(p+\epsilon_n) - \hat{Q}_{Y|X=1}(p-\epsilon_n) & \text{if } X = 1, \\ \hat{Q}_{Y|X=0}(p+\epsilon_n) - \hat{Q}_{Y|X=0}(p-\epsilon_n) & \text{if } X = 0, \end{cases}$$

and ϵ_n is a bandwidth parameter satisfying $\epsilon_n \xrightarrow{n\to\infty} 0$ (Koenker 2005). To avoid division by zero, a tolerance parameter is introduced in the event that $d\hat{Q}(p) = 0$.

An estimate of the density $\hat{f}_x$ and the sparsity $\hat{s}_x$ for all observations is obtained using the **Qtools**'s function `sparsity` (which is based on the code of `quantreg::summary.rq`). The argument `se` specifies the method used to calculate the standard errors (`iid`, `nid`, `ker`), while `hs` is a logical flag to choose the bandwidth ϵ_n (the Hall and Sheather's (1988) bandwidth if `TRUE` (default) or the Bofinger's (1975) bandwidth if `FALSE`). Estimated density and sparsity from the fitted model in Section 3.1 are given below.

```
R> sparsity(fit, se = "nid", hs = TRUE)

$density
             0.1       0.25        0.5       0.75        0.9
  [1,] 0.02670032 0.05192653 0.07497496 0.05192653 0.02670032
  [2,] 0.05340063 0.05192653 0.05997997 0.05192653 0.02136025
```

```
  [3,] 0.02670032 0.05192653 0.07497496 0.05192653 0.02670032

[...]

$sparsity
            0.1      0.25       0.5      0.75       0.9
  [1,] 37.45274 19.25798 13.33779 19.25798 37.45274
  [2,] 18.72637 19.25798 16.67223 19.25798 46.81593
  [3,] 37.45274 19.25798 13.33779 19.25798 37.45274

[...]

$bandwidth
[1] 0.05340063
```

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 (see Section 3.4). Model's 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 modeling. This approach applies to a subclass of linear models

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

and linear heteroscedastic models

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

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. See also Zhao (2000) for an examination of the asymptotic properties of the restricted QR estimator.

The package **Qtools** provides the functions `rrq`, `rrq.fit` and `rrq.wfit` which are, respectively, the 'restricted' analogous of `rq`, `rq.fit`, and `rq.wfit` 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** (Canty and Ripley 2014; Davison and Hinkley 1997). 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.

```
R> data(esterase)
R> # Fit standard quantile regression
R> fit.rq <- rq(Count ~ Esterase, data = esterase, tau = c(.1,.25,.5,.75,.9))
```

```
R> yhat1 <- fitted(fit.rq)
R> # Fit restricted quantile regression
R> fit.rrq <- rrq(Count ~ Esterase, data = esterase, tau = c(.1,.25,.5,.75,.9))
R> yhat2 <- fitted(fit.rrq)
```

The predicted 90th centile curve crosses the 50th and 75th curves at lower esterase concentrations (Figure 7). The crossing is removed in predictions based on RRQs.

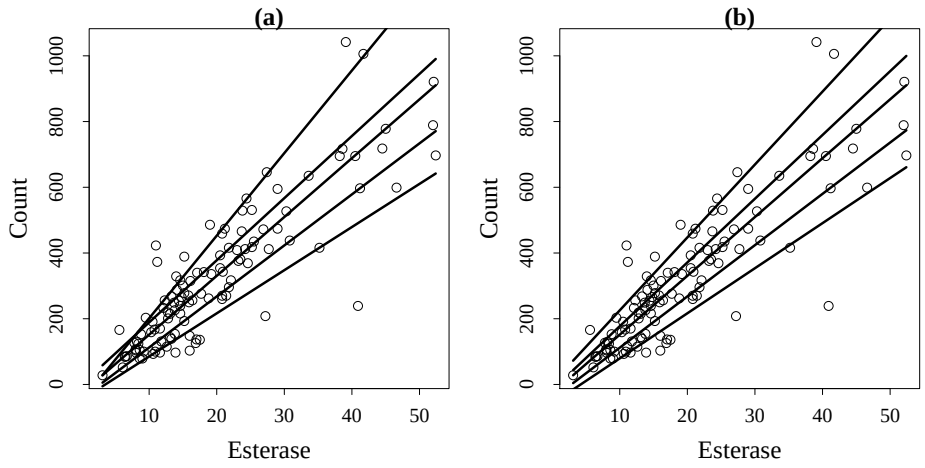

Figure 7: Predicted quantiles of number of bindings conditional on esterase concentration using regression quantiles (a) and restricted regression quantiles (b) in the Esterase data set.

For unconditional non-monotonic step functions $Q(p)$, monotonicity by rearrangement (Chernozhukov, Fernández-Val, and Galichon 2009) is provided by the function `quantreg::rearrange`.

3.3. Conditional LSS

Quantile-based measures of location, scale, and shape can be assessed *conditionally* on covariates. A simple approach to conditional LSS measures is to a fit a QR model of the type (8) and then predict the conditional quantiles according to specific values of $\mathbf{x}$. 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 analyzed here.

Suppose that the model of interest is

$$Q_{\text{ozone}}(p) = \beta_0(p) + \beta_1(p) \cdot \text{Solar.R.} \tag{11}$$

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

```
R> dd <- airquality[complete.cases(airquality),]
R> dd <- dd[order(dd$Solar.R),]
R> x <- seq(min(dd$Solar.R), max(dd$Solar.R), length = 200)
```

```
R> yhat <- predict(rq(Ozone ~ Solar.R , tau = c(.1,.5,.9), data = dd),
+ newdata = data.frame(Solar.R = x))
R> plot(Ozone ~ Solar.R, data = dd)
R> apply(yhat, 2, function(y,x) lines(x,y), x = x)
```

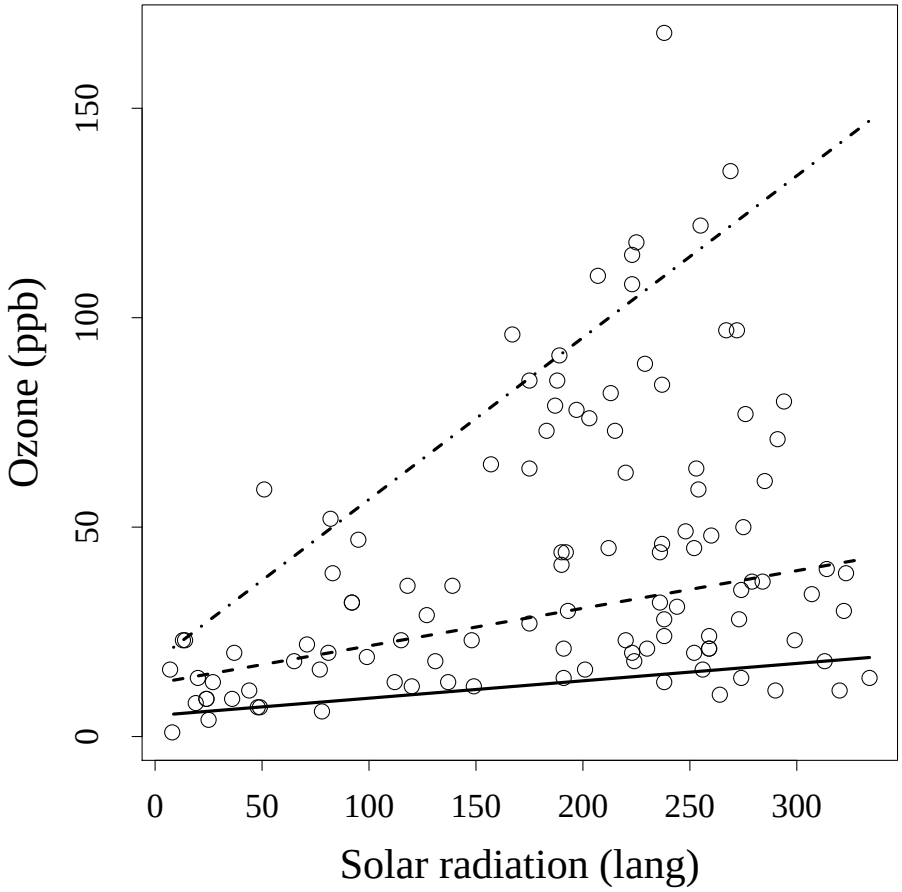

Figure 8: Predicted 10th (solid line), 50th (dashed line), and 90th (dot-dashed line) centiles of ozone conditional on solar radiation in the New York Air Quality data set.

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 8). The 90th centile of conditional ozone shows a steeper slope at 4.3 ppb/Ly, about nine times larger than the slope of the conditional 10th centile at 0.04 ppb/Ly. (The assumption that a straight-line model holds for these three conditional quantiles will be assessed in the next section.) Based on these results, it can be expected that solar radiation affects not only the location of the ozone distribution but also its scale and, possibly, its shape. The location-shift hypothesis for this model is rejected at the 10% level in favor of a more complex model. However, it seems that there is not enough evidence to support a model more complex than a location–scale-shift model.

```
R> kt <- KhmaladzeTest(Ozone ~ Solar.R, data = dd, taus = seq(.05,.95,by = .01),
```

```
+ nullH = "location")
R> KhmaladzeFormat(kt, 0.05)

Khmaladze test for the location-shift hypothesis
Joint test is significant at 10% level
Test(s) for individual slopes:
 significant at 10% level

R> kt <- KhmaladzeTest(Ozone ~ Solar.R, data = dd, taus = seq(.05,.95,by = .01),
R> nullH = "location-scale")
R> KhmaladzeFormat(kt, 0.05)

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

Inference on 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`. The argument `type` specifies the required type of regression model, more specifically `rq` for linear models and `rqt` for transformation-based models (see Section 4.1). As seen in Equation 4, the other probabilities of interest are $1 - p$, 0.25, 0.5, and 0.75, which are used by `qlss.formula` to obtain the decomposition of the conditional quantiles.

```
R> set.seed(567)
R> fit.qlss <- qlss(formula = Ozone ~ Solar.R, data = airquality, type = "rq",
+ probs = c(0.05, 0.1), predictLs = list(newdata = data.frame(Solar.R = x)),
+ ci = TRUE, R = 500)
R> str(fit.qlss)

List of 4
 $ location:List of 1

[...]

 $ scale   :List of 2

[...]

 $ shape   :List of 2

[...]

 $ CI      :List of 5

[...]

 - attr(*, "class")= chr "qlss"
```

The output, which is of class 'qlss', is a named list containing three elements as seen in the case of unconditional quantiles. However, the LSS measures of the distribution of daily mean ozone are now conditional on solar radiation. By default, the predictions are the fitted values. Alternatively, one can provide a new data frame via the argument `predictLs`, which consists of a list of arguments passed to the function `quantreg::predict.rq`. So, for example, the code above defines a fine grid of 300 values for solar radiation, i.e. `seq(min(x), max(x), length = 300)`. Finally, the function `qlss.formula` will take any additional argument to be passed to `quantreg::rq` (e.g., `subset`, `weights`, etc.).

The conditional LSS measures can be conveniently plotted using the `plot.qlss` function as shown in the code below.

```
R> plot(fit.qlss, z = x, which = 2, ci = TRUE, level = 0.90, type = "l",
+ xlab = "Solar radiation (lang)", lwd = 2)
```

The first argument specifies a `qlss.formula` object, while the second argument specifies the values of the covariate of interest against which the LSS measures must be plotted (note that these values must be the same as those used to predict the quantiles). Confidence intervals of a given level can be plotted by setting the argument `ci = TRUE` (note, however, that `ci` must be set to `TRUE` in the original `qlss` call as well). These are based on 500 bootstrap replications obtained with `quantreg::summary.rq` (it is, therefore, advisable to set the seed before calling `qlss`). Finally, the argument `which` specifies which of the probabilities given in `probs` should be used for plotting (in this example, `which = 2` corresponds to $p = 0.1$).

Figure 9 shows that both the median and the IQR of ozone increase with increasing solar radiation, as expected. The distribution of ozone is skewed to the right and the degree of asymmetry increases with increasing solar radiation. The conditional shape index increases monotonically from 1.16 to about 1.70, and it remains always below the tail-weight threshold of a normal distribution (1.90). The confidence intervals indicate a substantial estimation uncertainty in the proximity of the extremes of the observed solar radiation range.

3.4. 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 8 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. The sampling distribution

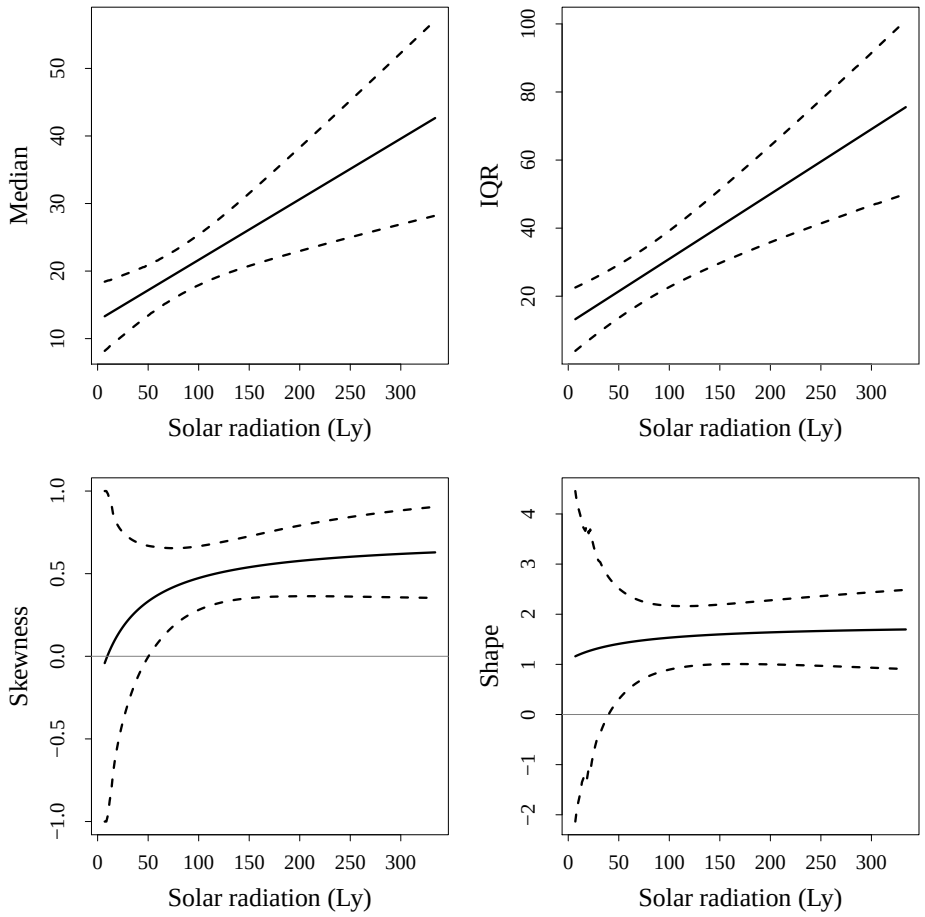

Figure 9: Location, scale and shape of ozone levels conditional on solar radiation in the New York Air Quality data set. Dashed lines denote the bootstrapped 90% point-wise confidence intervals.

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.

Turning to a practical example, the residual cusum test applied to the model in Equation 11 provides evidence of lack of fit for all quantiles considered, particularly for $p = 0.1$ and $p = 0.5$.

```
R> fit.rq <- rq(Ozone ~ Solar.R, tau = c(.1,.5,.9), data = dd)
R> gof.rq <- GOFTest(fit.rq, alpha = 0.05, B = 1000, seed = 987)
R> gof.rq

Goodness-of-fit test for quantile regression based on the cusum process
Quantile 0.1: Test statistic = 0.1057; p-value = 0.005
Quantile 0.5: Test statistic = 0.2191; p-value = 0
Quantile 0.9: Test statistic = 0.0457; p-value = 0.066
```

4. Other topics in conditional modeling

4.1. 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

$$Q_{Y|X}(p) = g\left\{\mathbf{x}^\top \boldsymbol{\beta}(p)\right\} \tag{12}$$

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

1. nonlinear parametric models: this approach may provide a model with 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;

2. polynomial models and smoothing splines: this approach goes 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;

3. 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, Cortina-Borja, and Geraci (2015) 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 transformation-based QR is to fit the model

$$Q_{h(Y;\lambda_p)}(p) = \mathbf{x}^\top \boldsymbol{\beta}(p). \tag{13}$$

The assumption is that the transformation h is the inverse of g, $h(Y;\lambda_p) \equiv g^{-1}(Y)$, so that the pth quantile function of the transformed response variable in Equation 12 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 (13) by virtue of the equivariance property (Q-transformation rule) of quantiles, i.e.

$$Q_{Y|X}(p) = h^{-1}\left\{\mathbf{x}^\top \boldsymbol{\beta}(p); \lambda_p\right\}. \tag{14}$$

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.

$$\frac{\partial Q(p)}{\partial x_j} = \frac{\partial Q(p)}{\partial \eta(p)} \cdot \frac{\partial \eta(p)}{\partial x_j}, \tag{15}$$

$\eta(p) = \mathbf{x}^\top \boldsymbol{\beta}(p)$. Once the estimates $\hat{\boldsymbol{\beta}}(p)$ and $\hat{\lambda}_p$ are obtained, these can be plugged in Equations 14 and 15.

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 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, $0 < y < 1$. (The symmetry here is that $h(Y;\lambda_p) = h(Y;-\lambda_p)$.) Geraci and Jones (2015) developed two families of transformations which can be applied to either singly or doubly bounded responses:

- Proposal I transformations – this family has one parameter and it comes in both symmetric and asymmetric forms;
- Proposal II transformations – this family has two parameters, with one parameter modeling 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 modeling. 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 recognized in QR modeling (Powell 1991; Buchinsky 1995; Chamberlain 1994; Mu and He 2007). Model 13 is the Box–Cox QR model if

$$h\left(Y;\lambda_p\right)=\begin{cases}\dfrac{Y^{\lambda_p}-1}{\lambda_p} & \text{if } \lambda_p\neq 0\\ \log Y & \text{if } \lambda_p=0.\end{cases} \tag{16}$$

Note that when $\lambda_p \neq 0$, the range of this transformation is not $\mathbb{R}$ but the singly bounded interval $(-1/\lambda_p, \infty)$. This implies that the inversion in (14) is defined only for $\lambda_p \mathbf{x}^\top \boldsymbol{\beta}(p)+1 > 0$. To overcome this computational difficulty, Geraci and Jones (2015) proposed to use instead

$$h\left(Y;\lambda_p\right)=\begin{cases}\dfrac{1}{2\lambda_p}\left(Y^{\lambda_p}-\dfrac{1}{Y^{\lambda_p}}\right) & \text{if } \lambda_p\neq 0\\ \log Y & \text{if } \lambda_p=0,\end{cases} \tag{17}$$

which has range $\mathbb{R}$ for all λ_p and hence 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. Note also that, for $\lambda_p \neq 0$, transformation (17) can be written

$$h\left(Y;\lambda_p\right)=\frac{1}{\lambda_p}\sinh(\lambda_p \log Y).$$

Model fitting for one-parameter transformation models (two-parameter transformation models are discussed further on in this paper) can be carried out using the function `tsrq`. The latter applies a two-stage (TS) estimator (Chamberlain 1994; Buchinsky 1995) whereby $\boldsymbol{\beta}(p)$ is estimated conditionally on a fine grid of values for λ_p. The `formula` argument specifies a linear model as in (13), while the argument `tsf` provides the desired transformation h. In **Qtools**, there are currently three one-parameter transformation families available: `mcjI` for proposal I transformations (Geraci and Jones 2015), `bc` for the Box–Cox model (Powell 1991), and `ao` for Aranda-Ordaz families (Aranda-Ordaz 1981). The reader is referred to the cited publications for details on these 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); and `lambda`, a numerical vector to define the grid of values for λ_p. An instance of the function `tsrq` based on (17) is given below

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

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

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 output 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 are obtained by plugging $\hat{\boldsymbol{\beta}}(p)$ and $\hat{\lambda}_p$ in (14) and (15), respectively. In particular, the function `maref` will determine the transformation h that was used for estimation and calculate the corresponding derivative. The argument `index` selects the position in the vector $\hat{\boldsymbol{\beta}}(p)$ of the effect of interest. Since the model may contain interactions, additional terms to be included in the computation can be specified in the argument `index.extra`.

```
R> x <- seq(min(dd$Solar.R), max(dd$Solar.R), length = 200)
R> yhat <- yhat <- predict(fit.rqt, newdata = data.frame(Solar.R = x),
+ type = "response")
R> dyhat <- maref(fit.rqt, newdata = data.frame(Solar.R = x), index = 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 (Figure 10). It might be worth testing the goodness-of-fit of the model. In Section 3.4, it was found evidence of lack of fit for the linear specification (11). In contrast, the output reported below indicates that, in general, the goodness of fit of the quantile models based on transformation (17) has improved since the test statistics are now smaller at all values of p. However, such improvement is still far from being sufficient for $p = 0.5$.

```
R> 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 = 0.077
Quantile 0.5: Test statistic = 0.1465; p-value = 0
Quantile 0.9: Test statistic = 0.0212; p-value = 0.289
```

There are other functions to fit transformation models. The function `rcrq` fits one-parameter transformation models using an estimator based on the RC process (akin to the RC process

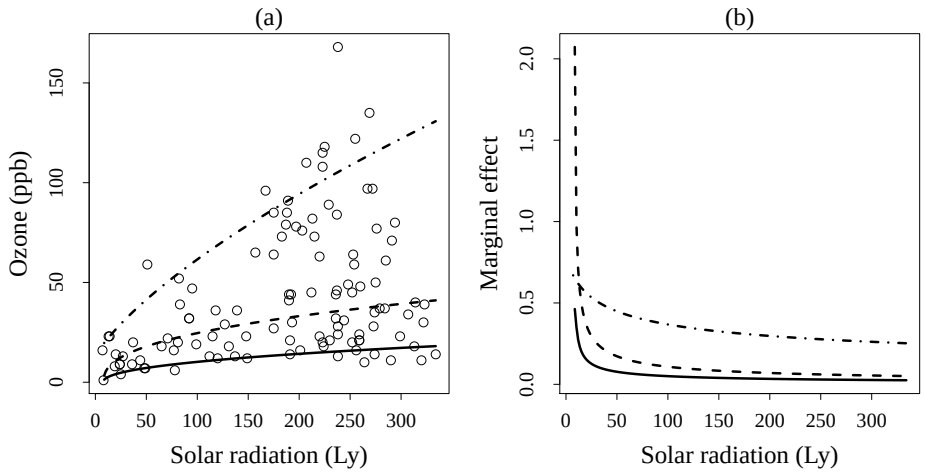

Figure 10: Predicted 10th (solid line), 50th (dashed line), and 90th (dot-dashed line) centiles of ozone conditional on solar radiation (a) and corresponding estimated marginal effects (b) using the symmetric proposal I transformation in the New York Air Quality data set.

introduced in Section 3.4) (Mu and He 2007). This estimator avoids the troublesome inversion of the Box-Cox and Aranda-Ordaz transformations, but it is computationally more intensive than the TS 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.

A summary of the basic differences between all fitting functions is given in Table 2. 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**. 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. A naïve approach to confidence interval estimation for $\boldsymbol{\beta}_p$ is to apply one of the several methods developed for linear quantile regression estimators (Koenker 2005, p.110) by assuming that the transformation parameter is known (see options `rank`, `iid`, `nid`, `ker`, and `boot` in `quantreg::summary.rq`). It is worth stressing that conditional inference may produce substantially lower standard errors for the regression coefficients and thus lead to overstating their significance (Mu and He 2007).

Function name	Transformation parameters	Estimation	Standard errors/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`	All types

Table 2: 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**.

4.2. 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 2013). Let the $n \times (k+1)$ matrix $\mathbf{Z}$ with row vectors $\mathbf{z}_i^\top$, $i = 1, \ldots, n$ collect the data in the study. Correspondingly, $\tilde{\mathbf{Z}}$ is used to denote the matrix without the jth column. In addition, for $j = 1, \ldots, k+1$, let n_j and $\bar{n}_j = n - n_j$ be, respectively, the number of observed and missing values in $\mathbf{z}_j$, and let A_j be the set indexing the units i for which the jth variable is not observed. The aim is to impute $\bar{n}_j$ missing values of a partially-observed continuous variable within a fully conditional specification (FCS) algorithm (van Buuren 2007). Throughout this section, it is assumed that the data are missing at random (MAR), i.e the probability of a value being missing conditional on observed data is independent of the unobserved data (Schafer 1997). Under this assumption, the application of MI is apposite.

The literature on multiple imputation (MI) is largely focused on location-shift models. For example, a common imputation model for continuous responses is the *iid* linear model, i.e. $Z_{i,j} \sim \mathrm{N}\left(\tilde{\mathbf{z}}_i^\top \beta, \sigma^2\right)$, $i \in A_j$. If the normality assumption is too restrictive, an alternative approach is to draw a sample from $F_{Z_{i,j}|\tilde{Z}}$ using the probability integral transformation. The latter is a well-known theorem used in random number generation: if $V \sim F$ and $U \sim \mathrm{Unif}\,(0,1)$, then $F^{-1}(U) \sim F$, that is V and $F^{-1}(U)$ have the same distribution. Thus, a distribution-free approach to imputation can be carried out in three steps:

(a) Generate u independently from a standard uniform distribution. To avoid sampling in the vicinity of the boundaries which could cause computational inconveniences in step (b), the sampling domain can be restricted to $\mathrm{Unif}(\omega, 1-\omega)$ with ω sufficiently small, say $\omega = 0.001$. The parameter ω can be used to trim or truncate the distribution of $Z_{i,j}$.

(b) Estimate the quantile regression model

$$Q_{Z_j|\tilde{Z}}(u) = \tilde{\mathbf{z}}^\top \boldsymbol{\beta}(u).$$

(c) Obtain and impute the value $z^*_{i,j} = \tilde{\mathbf{z}}_i^\top \hat{\boldsymbol{\beta}}(u)$, $i \in A_j$.

The imputation procedure above can be easily extended to models fitted on the transformed variable $h(Z_j)$, for some monotone function h, in which case the imputed value would be $z^*_{i,j} = h^{-1}\left\{\tilde{\mathbf{z}}_i^\top \hat{\boldsymbol{\beta}}(u)\right\}$. The equivariance property is therefore very useful if a transformation is applied to achieve linearity of the conditional model or to ensure that imputations lie within some interval (a, b). Some authors refer to such pre-imputation transformations as pre-processing, followed by post-processing to transform the data back (Su, Gelman, Hill, and Yajima 2011).

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 described in Section 4.1. An example available from `?mice.impute.rq` using the `nhanes` data set is reported below.

```
R> require(mice)
R> data(nhanes)
R> nhanes2 <- nhanes
R> nhanes2$hyp <- as.factor(nhanes2$hyp)
R> # Impute continuous variables using quantile regression
R> set.seed(199)
R> imp <- mice(nhanes2, meth = c("polyreg", "rq", "logreg", "rq"), m = 5)
R> # estimate linear regression and pool results
R> fit <- lm.mids(bmi ~ hyp + chl, data = imp)
R> pool(fit)

Call: pool(object = fit)

Pooled coefficients:
(Intercept)        hyp2         chl
23.09613686  0.79895370  0.01652075

Fraction of information about the coefficients missing due to nonresponse:
(Intercept)        hyp2         chl
  0.3626414   0.6585308   0.4835364
```

4.3. Conditional quantiles of discrete data

Modeling discrete response variables, such as categorical and count responses, has been traditionally approached with distribution-based methods: a parametric model $F_{Y|X}(y;\theta)$ is assumed and then fitted by means of MLE. Binomial, negative binomial, multinomial and Poisson regressions are well-known in many applied sciences. Because of the computational advantages and the asymptotic properties of MLE, these methods have long ruled among competing alternatives.

Modeling 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.

Let Y be a count variable such as, for example, the number of car accidents during a week or the number of visits of a patient to their doctor during a year. As usual, X denotes a vector of covariates. Poisson regression, which belongs to the family of generalized 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

$$\theta \equiv \mathsf{E}(Y|X=\mathbf{x}) = h^{-1}\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$$

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

$$Q_{h(Z;p)}(p) = \mathbf{x}^\top \boldsymbol{\beta}(p). \tag{18}$$

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. A natural choice for count data is a log-linear model (Machado and Santos Silva 2005), i.e.

$$h(Z;p) = \begin{cases} \log(Z - p) & \text{for } Z > p \\ \log \zeta & \text{for } Z \leq p. \end{cases}$$

where $0 < \zeta < p$. It follows that $Q_{Z|X}(p) = p + \exp\left(\mathbf{x}^\top \boldsymbol{\beta}(p)\right)$. (Note that the pth quantile of the conditional distribution of Z is bounded below by p.) Given the continuity between counts induced by jittering, standard inference for linear quantile functions (Koenker and Bassett 1978) can be applied to fit (18). 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 (18), 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 described in Section 4.1 are allowed. In the example below, estimation is carried out using $M = 50$ jittered samples and $\zeta = 10^{-5}$ (see Machado and Santos Silva (2005) for further details on these settings).

```
R> data(esterase)
R> fit.rq.counts <- rq.counts(formula = Count ~ Esterase, tau = 0.1,
+ data = esterase, tsf = "bc", lambda = 0, M = 50, zeta = 1e-05)
```

Figure 11 shows a contrast between centile curves as predicted by the Poisson and the QR models in the Esterase data set. The Poisson distribution clearly underestimates the variability in the data. An empirical modeling of the conditional quantiles seems to be preferred in this case. Of course, the assumption of log-linearity of the models would need to be carefully assessed (note that `GOFTest` can be applied also to `rq.counts` objects).

Qtools provides functions for modeling 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 $\mathbf{x}$, that is

$$\pi \equiv \mathsf{E}(Y|X = \mathbf{x}) = h^{-1}\left(\mathbf{x}^\top \boldsymbol{\beta}\right).$$

Common choices for the link function $h : (0, 1) \to \mathbb{R}$ are logit, probit and c-log-log, which can be considered special cases of the Aranda-Ordaz families of transformations. Another way to formulate the regression problem above is to consider the continuous latent variable Y^*, where

$$Y^* = \mathbf{x}^\top \boldsymbol{\beta} + \epsilon, \tag{19}$$

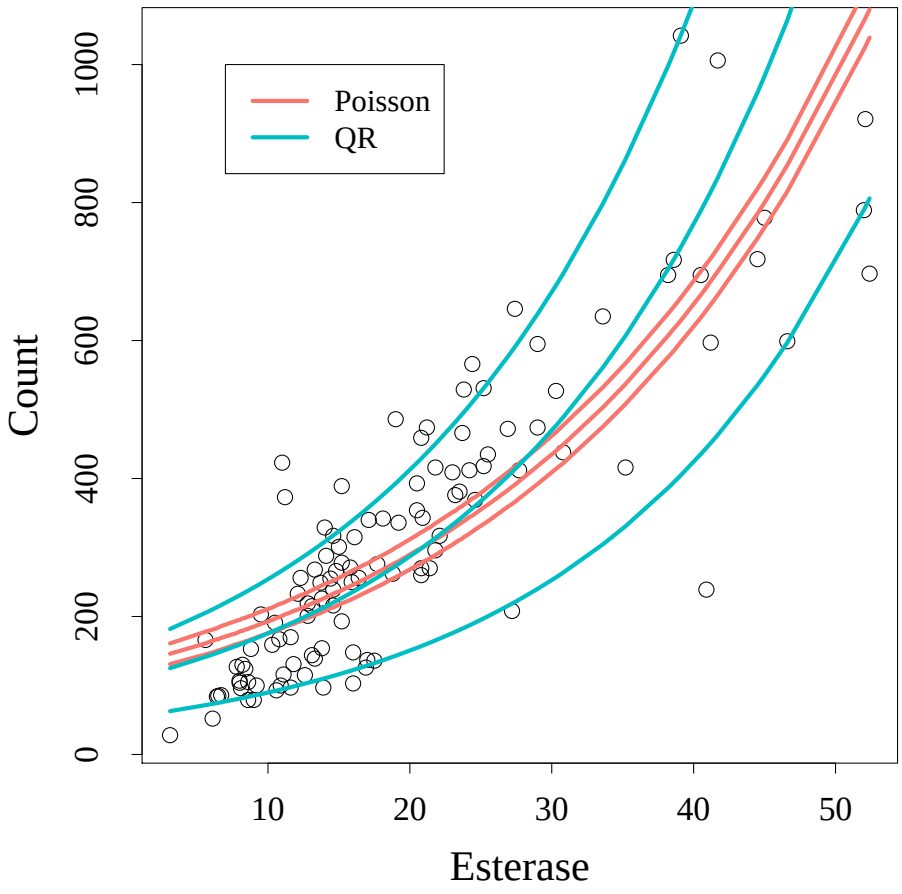

Figure 11: Predicted 10th, 50th, and 90th centiles of number of bindings conditional on esterase concentration using Poisson regression and distribution-free quantile regression (QR) in the Esterase data set.

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

$$\begin{aligned}\Pr(Y = 1) &= \Pr(Y^* > 0) \\ &= \Pr\left(\mathbf{x}^\top \boldsymbol{\beta} + \epsilon > 0\right) \\ &= \Pr\left(\epsilon > -\mathbf{x}^\top \boldsymbol{\beta}\right) \\ &= 1 - F_\epsilon\left(-\mathbf{x}^\top \boldsymbol{\beta}\right),\end{aligned}$$

with the understanding that the probabilities above are conditional on $\mathbf{x}$. For distributions symmetric about 0, the expression above can be re-written as $\pi = F_\varepsilon\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$, where $\varepsilon = -\epsilon$. Parametric models for F_ε will correspond to specific link functions $h \equiv F_\varepsilon^{-1}$. For example, the probit function corresponds to $\varepsilon \sim \mathrm{N}(0, 1)$. In the econometric literature, this formulation is known as binary choice model.

Maximum score estimation, originally developed by Manski (1975, 1985), is equivalent to estimating the conditional quantiles of the latent variable Y^* in (19). Using the same notation introduced in Section 3.2, the problem to be solved is given by

$$\min_{b \in \mathbb{R}^k} \sum_{i=1}^{n} \kappa_p\left(y_i - I\left(\mathbf{x}_i^\top \mathbf{b} > 0\right)\right). \tag{20}$$

Since the indicator function is a monotone transformation, the Q-transformation rule applies. Indeed, $I\left(Q_{Y^*|X}(p) > 0\right) = Q_{I(Y^*)>0|X}(p) = Q_{Y|X}(p)$. In practice, the goal is to find $\mathbf{b}$ such that the number of matches between observed and fitted 0's and 1's is as large as possible. However, the minimization problem in (20) offers numerical challenges due to the the piecewise linearity of the indicator function and the nonconvexity of the loss function. Smoothed approximations of $I(\cdot)$ and simulated annealing algorithms have been suggested (Horowitz 1992; Kordas 2006).

It should be noted that the regression parameter $\boldsymbol{\beta}(p)$ is identified up to a scale. For this reason, a normalization is required. In one approach (Horowitz 1992), it is assumed that there exists a regressor, say x_k, such that, conditionally on the remaining terms

1. the probability distribution of x_k is absolutely continuous (continuity assumption);
2. the distribution of ϵ is conditionally independent from x_k (homoscedasticity assumption).

If the latter assumption does not hold, 'slopes' are no longer comparable across different quantile models (Kordas 2006). This issue can be addressed using the normalization $||\boldsymbol{\beta}(p)|| = 1$ (Manski 1975, 1985), although, in this case, a meaningful interpretation of the intercepts would be lost.

Let us consider the following data set on wine quality (Cortez, Cerdeira, Almeida, Matos, and Reis 2009). The data set consists of 1599 observations on 12 variables (11 physicochemical continuous attributes and one sensory categorical variable) for red variants of the Portuguese 'Vinho Verde' – white wines are excluded from the present analysis. The outcome of interest is 'quality' of the wine as assessed on a 1-10 scale, with 1 and 10 indicating, respectively, worst and best quality. The summary of the data set is given below.

```
R> wine <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/
wine-quality/winequality-red.csv", sep = ";")
R> summary(wine)

 fixed.acidity   volatile.acidity  citric.acid    residual.sugar
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
       pH          sulphates         alcohol         quality
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

The rankings of the red wines range from 3 to 8 and are almost equally split at the mid-ranking 5.5.

```
> table(wine$quality)

  3   4   5   6   7   8
 10  53 681 638 199  18
```

Given the limited number of unique rankings and to simplify the analysis, a logistic regression is fitted on the binary variable `y` which denotes whether a wine is above (`y = 1`) or below (`y = 0`) the mid-rank. All the physicochemical variables enter in the model as covariates (issues of multicollinearity are not discussed here).

```
R> wine$y <- as.numeric(wine$quality > 5)
R> ff <- as.formula(y ~ fixed.acidity + volatile.acidity + citric.acid +
+ residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
+ density + sulphates + alcohol + pH)
R> summary(glm(ff, data = wine, family = binomial("logit")))

Call:
glm(formula = ff, family = binomial("logit"), data = wine)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4025  -0.8387   0.3105   0.8300   2.3142

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           42.949948  79.473979   0.540  0.58890
fixed.acidity          0.135980   0.098483   1.381  0.16736
volatile.acidity      -3.281694   0.488214  -6.722 1.79e-11 ***
citric.acid           -1.274347   0.562730  -2.265  0.02354 *
residual.sugar         0.055326   0.053770   1.029  0.30351
chlorides             -3.915713   1.569298  -2.495  0.01259 *
free.sulfur.dioxide    0.022220   0.008236   2.698  0.00698 **
total.sulfur.dioxide  -0.016394   0.002882  -5.688 1.29e-08 ***
density              -50.932385  81.148745  -0.628  0.53024
sulphates              2.795107   0.452184   6.181 6.36e-10 ***
alcohol                0.866822   0.104190   8.320  < 2e-16 ***
pH                    -0.380608   0.720203  -0.528  0.59717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[...]
```

The estimated coefficients for volatile acidity ($g \cdot dm^{-3}$), citric acid ($g \cdot dm^{-3}$), chlorides ($g \cdot dm^{-3}$), and total sulfur dioxide ($mg \cdot dm^{-3}$) show a negative association with the probability of a wine being ranked above 'mediocrity'. The association is positive for free sulfur dioxide ($mg \cdot dm^{-3}$), sulphates ($g \cdot dm^{-3}$), and alcohol (% vol.). The other variables are not significant at the 5% level.

Let us now consider a binary quantile regression approach to model the (latent) quality of the wines. 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, Ferrier, and Rogers 1994). **Qtools** offers general methods for objects of class `rq.bin` including `coef` and `predict`. In particular, the commands `predict(object, type = "latent")` and `predict(object, type = "probability")` provide predictions for, respectively, the latent quantiles $\mathbf{x}_i^\top \boldsymbol{\beta}(p)$ and the individual probabilities $\Pr(Y_i^* > 0)$, $i = 1, \ldots, n$. A `summary` method to calculate standard errors has not yet been implemented.

Once the linear model has been defined, it is important to decide on the normalization approach. Horowitz' approach, the default, is specified with the argument `normalize = "last"` in `rq.bin`. The user must ensure that the last term in `formula` (or the last column in the matrix `x` when using `rqbin.fit`) corresponds to the regressor x_k discussed above. Manski's approach is obtained with `normalize = "all"`. In the Wine Quality data set, the continuity and homoscedasticity assumptions for pH seem to be reasonably met. This can be verified also by analyzing the residuals of the model without pH as shown below.

```
R> tmp <- glm(y ~ fixed.acidity + volatile.acidity + citric.acid +
+ residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
```

```
+ density + sulphates + alcohol, data = wine, family = binomial())
R> w <- residuals(tmp, type = "deviance")
R> summary(lm(wine$pH ~ w))
R> summary(rq(wine$pH ~ w, tau = 1:9/10), se = "nid")
```

The choice of the quantiles to be estimated should take into account the probability of the event in the population to ensure that both positive and negative values of $\mathbf{x}^\top \boldsymbol{\beta}(p)$ appear in the population (Manski 1985). Let α be such probability. If α is small (rare event), then one may consider estimating the quantile $p = 1 - \alpha$.

Given the proportion of cases in the sample is about 50%, four binary regressions are fitted for the first four deciles $p \in \{0.1, 0.2, 0.3, 0.4\}$ to investigate how covariates affect the latent quality of wines with probability below or near $\Pr(Y^* > 0)$.

```
R> fit.bin.rq <- rq.bin(ff, tau = 1:4/10, data = wine, normalize = "last")
R> fit.bin.rq

Call: rq.bin(formula = ff, tau = 1:4/10, data = wine, normalize = "last")

Binary quantile model

Coefficients (last coefficient is set equal to 1):
                      tau = 0.1     tau = 0.2     tau = 0.3     tau = 0.4
(Intercept)         -59.71800365  -75.41198294  -77.09631530  -78.01551696
fixed.acidity        -1.79404630    0.54662796   -6.77954027    5.84083844
volatile.acidity     -0.02877108    5.29573448   -2.87345046   -2.45791298
citric.acid          12.02411499    4.75218615   12.12135441   10.99106011
residual.sugar        1.32124832    3.20378316   -5.12604687    0.63582415
chlorides            16.68480346   16.73069252   10.23576481   14.36408494
free.sulfur.dioxide   6.98601986    8.09025703    6.03302707    6.95053071
total.sulfur.dioxide -5.68934652   -6.58013730   -4.72346519   -6.66718453
density              92.08799058   79.09454731   93.62448607   79.37356139
sulphates            -1.09200785    3.78027756    3.83931880   -2.83597864
alcohol              -1.57289484   -1.11292830    5.40091225    7.45298172
pH                    1.00000000    1.00000000    1.00000000    1.00000000

Degrees of freedom: 1599 total; 1588 residual
```

The output above indicates that the association between some of the covariates and quality may be heterogeneous across the quantiles of the latent variable. For example, wines with higher alcohol percentages are perceived as of better quality as long as they rank in the third or fourth decile of the conditional distribution. However, at lower quantiles, alcohol seems not to affect, or to affect negatively, perceived wine quality. In other words, for wines that rank poorly as compared to other wines with similar attributes, a higher alcohol content will not improve their perceived quality conditional on those attributes.

5. 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 modeling.

Acknowledgements

This work is partially supported by an ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina. The maintainers of the UCI Machine Learning Repository (Lichman 2013) are gratefully acknowledged for making the Wine Quality data set available.

References

Aranda-Ordaz FJ (1981). "On Two Families of Transformations to Additivity for Binary Response Data." *Biometrika*, **68**(2), 357–363.

Azzalini A, Bowman AW (1990). "A Look at Some Data on the Old Faithful Geyser." *Journal of the Royal Statistical Society C*, **39**(3), 357–365.

Bassett G, Koenker R (1978). "Asymptotic Theory of Least Absolute Error Regression." *Journal of the American Statistical Association*, **73**(363), 618–622.

Benjamini Y, Krieger AM (1996). "Concepts and Measures for Skewness with Data-Analytic Implications." *Canadian Journal of Statistics*, **24**(1), 131–140.

Bofinger E (1975). "Estimation of a Density Function Using Order Statistics." *Australian Journal of Statistics*, **17**(1), 1–7.

Bottai M, Zhen H (2013). "Multiple Imputation Based on Conditional Quantile Estimation." *Epidemiology, Biostatistics, and Public Health*, **10**(1), e8758.

Box GEP, Cox DR (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society B*, **26**(2), 211–252.

Buchinsky M (1995). "Quantile Regression, Box-Cox Transformation Model, and the US Wage Structure, 1963-1987." *Journal of Econometrics*, **65**(1), 109–154.

Canty A, Ripley BD (2014). ***boot**: Bootstrap R (S-Plus) Functions*. R package version 1.3-15, URL http://CRAN.R-project.org/package=boot.

Chamberlain G (1994). *Quantile Regression, Censoring, and the Structure of Wages*, volume 1. Cambridge University Press, Cambridge, UK.

Chernozhukov V, Fernández-Val I, Galichon A (2009). "Improving Point and Interval Estimators of Monotone Functions by Rearrangement." *Biometrika*, **96**(3), 559–575. doi:10.1093/biomet/asp030. URL http://biomet.oxfordjournals.org/content/96/3/559.abstract.

Cortez P, Cerdeira A, Almeida F, Matos T, Reis J (2009). "Modeling Wine Preferences by Data Mining From Physicochemical Properties." *Decision Support Systems*, **47**(4), 547–553.

David HA (1995). "First (?) Occurrence of Common Terms in Mathematical Statistics." *The American Statistician*, **49**(2), 121–133.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.

Dehbi HM, Cortina-Borja M, Geraci M (2015). "Aranda-Ordaz Quantile Regression for Student Performance Assessment." *Journal of Applied Statistics (in press)*.

Doksum K (1974). "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case." *The Annals of Statistics*, **2**(2), 267–277.

Galton F (1875). "IV. Statistics by Intercomparison, with Remarks on the Law of Frequency of Error." *Philosophical Magazine Series 4*, **49**(322), 33–46.

Galton F (1882). "Report of the Anthropometric Committee." In *Report of the 51st Meeting of the British Association for the Advancement of Science, 1881*, pp. 245–260.

Galton F (1885). "Some Results of the Anthropometric Laboratory." *The Journal of the Anthropological Institute of Great Britain and Ireland*, **14**, 275–287.

Geraci M (2013). "Estimation of Regression Quantiles in Complex Surveys with data Missing at Random: An Application to Birthweight Determinants." *Statistical Methods in Medical Research*. doi:10.1177/0962280213484401.

Geraci M, Jones MC (2015). "Improved Transformation-Based Quantile Regression." *Canadian Journal of Statistics*, **43**(1), 118–132.

Gilchrist W (2000). *Statistical Modelling with Quantile Functions*. Chapman & Hall/CRC, Boca Raton, FL.

Goffe WL, Ferrier GD, Rogers J (1994). "Global Optimization of Statistical Functions With Simulated Annealing." *Journal of Econometrics*, **60**(1), 65–99.

Groeneveld RA (1998). "A Class of Quantile Measures for Kurtosis." *The American Statistician*, **52**(4), pp. 325–329.

Groeneveld RA, Meeden G (1984). "Measuring Skewness and Kurtosis." *Journal of the Royal Statistical Society D*, **33**(4), pp. 391–399.

Hald A (1998). *A History of Mathematical Statistics from 1750 to 1930*. John Wiley & Sons, New York, NY.

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

Hall P, Sheather SJ (1988). "On the Distribution of a Studentized Quantile." *Journal of the Royal Statistical Society Series B*, **50**(3), 381–391.

He X (1997). "Quantile Curves without Crossing." *The American Statistician*, **51**(2), 186–192.

He XM, Zhu LX (2003). "A Lack-of-fit Test for Quantile Regression." *Journal of the American Statistical Association*, **98**(464), 1013–1022.

Horn PS (1983). "A Measure for Peakedness." *The American Statistician*, **37**(1), 55–56.

Horowitz JL (1992). "A Smoothed Maximum Score Estimator For the Binary Response Model." *Econometrica*, **60**(3), 505–531.

Hyndman RJ, Fan Y (1996). "Sample Quantiles in Statistical Packages." *The American Statistician*, **50**(4), 361–365.

Jones MC (2007). "Connecting Distributions with Power Tails on the Real Line, the Half Line and the Interval." *International Statistical Review*, **75**(1), 58–69.

Jones MC, Rosco JF, Pewsey A (2011). "Skewness-invariant Measures of Kurtosis." *The American Statistician*, **65**(2), 89–95.

Kendall MG (1940). "Note on the Distribution of Quantiles for Large Samples." *Supplement to the Journal of the Royal Statistical Society*, **7**(1), 83–85.

Khmaladze EV, Koul HL (2004). "Martingale Transforms Goodness-of-fit Tests in Regression Models." pp. 995–1034.

Koenker R (2005). *Quantile Regression*. Cambridge University Press, New York, NY.

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

Koenker R, Bassett G (1978). "Regression Quantiles." *Econometrica*, **46**(1), 33–50.

Koenker R, Bassett G (1985). "On Boscovich's Estimator." *The Annals of Statistics*, **13**(4), 1625–1628.

Koenker R, Machado JAF (1999). "Goodness of Fit and Related Inference Processes for Quantile Regression." *Journal of the American Statistical Association*, **94**(448), 1296–1310.

Koenker R, Park BJ (1996). "An Interior Point Algorithm for Nonlinear Quantile Regression." *Journal of Econometrics*, **71**(1-2), 265–283.

Koenker R, Xiao ZJ (2002). "Inference on the Quantile Regression Process." *Econometrica*, **70**(4), 1583–1612.

Kordas G (2006). "Smoothed Binary Regression Quantiles." *Journal of Applied Econometrics*, **21**(3), 387–407.

Lehmann EL (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, CA.

Lichman M (2013). "UCI Machine Learning Repository." URL http://archive.ics.uci.edu/ml.

Ma Y, Genton MG, Parzen E (2011). "Asymptotic Properties of Sample Quantiles of Discrete Distributions." *Annals of the Institute of Statistical Mathematics*, **63**(2), 227–243.

Machado JAF, Mata J (2000). "Box-Cox Quantile Regression and the Distribution of Firm Sizes." *Journal of Applied Econometrics*, **15**(3), 253–274.

Machado JAF, Santos Silva JMC (2005). "Quantiles for Counts." *Journal of the American Statistical Association*, **100**(472), 1226–1237.

Manski CF (1975). "Maximum Score Estimation of the Stochastic Utility Model of Choice." *Journal of Econometrics*, **3**(3), 205–228.

Manski CF (1985). "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator." *Journal of Econometrics*, **27**(3), 313–333.

Mu YM, He XM (2007). "Power Transformation Toward a Linear Regression Quantile." *Journal of the American Statistical Association*, **102**(477), 269–279.

Muñoz JF, Rueda M (2009). "New Imputation Methods for Missing Data Using Quantiles." *Journal of Computational and Applied Mathematics*, **232**(2), 305–317.

Oberhofer W, Haupt H (2005). "The Asymptotic Distribution of the Unconditional Quantile Estimator Under Dependence." *Statistics & Probability Letters*, **73**(3), 243–250.

Parente PMDC, Santos Silva JMC (2015). "Quantile Regression With Clustered Data." *Journal of Econometric Methods*. doi:10.1515/jem-2014-0011. Forthcoming.

Parzen E (1979). "Nonparametric Statistical Data Modeling." *Journal of the American Statistical Association*, **74**(365), 105–121.

Parzen E (2004). "Quantile Probability and Statistical Data Modeling." *Statistical Science*, **19**(4), 652–662.

Powell JL (1991). *Estimation of Monotonic Regression Models Under Quantile Restrictions*, pp. 357–384. Cambridge University Press, New York, NY.

Powell JL (1994). *Estimation of Semiparametric Models*, volume Volume 4, chapter 41, pp. 2443–2521. Elsevier.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ruppert D (1987). "What is Kurtosis?: An Influence Function Approach." *The American Statistician*, **41**(1), 1–5.

Sakia RM (1992). "The Box-Cox Transformation Technique: A Review." *Journal of the Royal Statistical Society D*, **41**(2), 169–178.

Schafer JL (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Staudte RG (2014). "Inference for Quantile Measures of Kurtosis, Peakedness and Tailweight." *arXiv preprint arXiv:1047.6461v1 [math.ST]*. URL http://arxiv.org/pdf/1407.6461v1.pdf.

Su YS, Gelman A, Hill J, Yajima M (2011). "Multiple Imputation With Diagnostics (mi) in R: Opening Windows Into the Black Box." *Journal of Statistical Software*, **45**(2), 1–31.

Tukey JW (1965). "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.

van Buuren S (2007). "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." *Statistical Methods in Medical Research*, **16**(3), 219–242.

van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1–67.

Wang JD (1995). "Asymptotic Normality of L_1-Estimators in Nonlinear Regression." *Journal of Multivariate Analysis*, **54**(2), 227–238.

Yu KM, Jones MC (1998). "Local Linear Quantile Regression." *Journal of the American Statistical Association*, **93**(441), 228–237.

Zhao QS (2000). "Restricted Regression Quantiles." *Journal of Multivariate Analysis*, **72**(1), 78–99.

Zheng JX (1998). "A Consistent Nonparametric Test of Parametric Regression Models under Conditional Quantile Restrictions." *Econometric Theory*, **14**(1), 123–138.

Affiliation:

Marco Geraci
Department of Epidemiology and Biostatistics
Arnold School of Public Health
University of South Carolina
915 Greene Street, Columbia SC 29208, USA
E-mail: geraci@mailbox.sc.edu
URL: http://marcogeraci.wordpress.com/