A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 12

Quantile Regression: Head Circumference for Age

12.1 Introduction

12.2 Quantile Regression

12.3 Analysis Using R

We begin with a graphical inspection of the influence of age on head circumference by means of a scatterplot. Plotting all pairs of age and head circumference in one panel gives more weight to the teens and 20s, so we produce one plot for younger boys between two and nine years old and one additional plot for boys older than nine years (or > 108 months, to be precise). The cut function is very convenient for constructing a factor representing these two groups

```
R> summary(db)
```

```
      head            age       
 Min.   :33.5   Min.   : 0.03  
 1st Qu.:48.8   1st Qu.: 1.75  
 Median :53.0   Median : 9.99  
 Mean   :51.7   Mean   : 8.94  
 3rd Qu.:55.7   3rd Qu.:14.84  
 Max.   :66.3   Max.   :21.68  
```

```
R> db$cut <- cut(db$age, breaks = c(2, 9, 23),
+                labels = c("2-9 yrs", "9-23 yrs"))
```

which can then be used as a conditioning variable for conditional scatterplots produced with the xyplot function (Sarkar, 2014, package **lattice**). Because we draw 5101 points in total, we use transparent shading (via rgb(.1, .1, .1, .1)) in order to obtain a clearer picture for the more populated areas in the plot.

Figure 12.1, as expected, shows that head circumference increases with age. It also shows that there is considerable variation and also quite a number of extremely large or small head circumferences in the respective age cohorts. It should be noted that each point corresponds to one boy participating in the study due to its cross-sectional study design. No longitudinal measurements (cf. Chapter **??**) were taken and we can safely assume independence between observations.

We start with a simple linear model, computed separately for the younger and older boys, for regressing the mean head circumference on age

```
R> (lm2.9 <- lm(head ~ age, data = db, subset = age < 9))
```

```
R> db$cut <- cut(db$age, breaks = c(2, 9, 23),
+                  labels = c("2-9 yrs", "9-23 yrs"))
R> xyplot(head ~ age | cut, data = db, xlab = "Age (years)",
+         ylab = "Head circumference (cm)",
+         scales = list(x = list(relation = "free")),
+         layout = c(2, 1), pch = 19,
+         col = rgb(.1, .1, .1, .1))
```

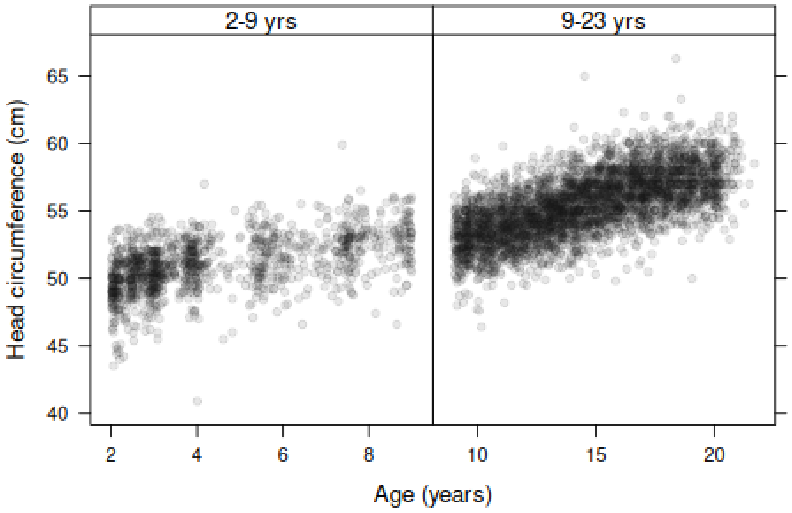

Figure 12.1 Scatterplot of age and head circumference for 5101 Dutch boys.

```
Call:
lm(formula = head ~ age, data = db, subset = age < 9)

Coefficients:
(Intercept)          age
      43.72         1.52
```

```
R> (lm9.23 <- lm(head ~ age, data = db, subset = age > 9))
```

```
Call:
lm(formula = head ~ age, data = db, subset = age > 9)

Coefficients:
(Intercept)          age
     48.619        0.469
```

This approach is equivalent to fitting two intercepts and two slopes in the joint model

```
R> (lm_mod <- lm(head ~ age:I(age < 9) + I(age < 9) - 1,
+                  data = db))
```

```
Call:
lm(formula = head ~ age:I(age < 9) + I(age < 9) - 1, data = db)
```

```
Coefficients:
    I(age < 9)FALSE       I(age < 9)TRUE  age:I(age < 9)FALSE
             48.620               43.715                0.469
 age:I(age < 9)TRUE
              1.517
```

while omitting the global intercept. Because the median of the normal distribution is equal to its mean, the two models can be interpreted as conditional median models under the normal assumption. The model states that within one year, the head circumference increases by 1.517 cm for boys less than nine years old and by 0.469 for older boys.

We now relax this distributional assumption and compute a median regression model using the `rq` function from package **quantreg** (Koenker, 2013):

```
R> library("quantreg")
R> (rq_med2.9 <- rq(head ˜ age, data = db, tau = 0.5,
+                    subset = age < 9))
```

```
Call:
rq(formula = head ~ age, tau = 0.5, data = db, subset = age <
    9)

Coefficients:
(Intercept)         age
      45.01        1.28

Degrees of freedom: 3193 total; 3191 residual
```

```
R> (rq_med9.23 <- rq(head ˜ age, data = db, tau = 0.5,
+                     subset = age > 9))
```

```
Call:
rq(formula = head ~ age, tau = 0.5, data = db, subset = age >
    9)

Coefficients:
(Intercept)         age
     48.579       0.472

Degrees of freedom: 3842 total; 3840 residual
```

When we construct confidence intervals for the intercept and slope parameters from both models for the younger boys

```
R> cbind(coef(lm2.9)[1], confint(lm2.9, parm = "(Intercept)"))
```

```
                 2.5 % 97.5 %
(Intercept) 43.7  43.6   43.9
```

```
R> cbind(coef(lm2.9)[2], confint(lm2.9, parm = "age"))
```

```
         2.5 % 97.5 %
age 1.52  1.47   1.57
```

```
R> summary(rq_med2.9, se = "rank")
```

```
Call: rq(formula = head ~ age, tau = 0.5, data = db, subset = age <
    9)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 45.01        44.81    45.21
age          1.28         1.23     1.38
```

we see that the two intercepts are almost identical but there seems to be a larger slope parameter for age in the median regression model. For the older boys, we get the confidence intervals via

```
R> cbind(coef(lm9.23)[1], confint(lm9.23, parm = "(Intercept)"))
```

```
                 2.5 % 97.5 %
(Intercept) 48.6  48.4   48.9
```

```
R> cbind(coef(lm9.23)[2], confint(lm9.23, parm = "age"))
```

```
          2.5 % 97.5 %
age 0.469 0.452  0.486
```

```
R> summary(rq_med9.23, se = "rank")
```

```
Call: rq(formula = head ~ age, tau = 0.5, data = db, subset = age >
    9)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 48.579       48.391   48.893
age          0.472        0.430    0.486
```

with again almost identical intercepts and only a slightly increased slope for age in the median regression model.

Since one of our aims was the construction of growth curves, we first use the linear models regressing head circumference on age to plot such curves. Based on the two normal linear models, we can compute the quantiles of head circumference for age. For the following values of τ

```
R> tau <- c(.01, .1, .25, .5, .75, .9, .99)
```

and a grid of age values

```
R> gage <- c(2:9, 9:23)
R> i <- 1:8
```

(the index `i` denoting younger boys), we compute the standard prediction intervals taking the randomness of the estimated intercept, slope, and variance parameters into account. We first set up a data frame with our grid of age values and then use the `predict` function for a linear model to compute prediction intervals, here with a coverage of 50%. The lower limit of such a 50% prediction interval is equivalent to the conditional 25% quantile for the given age and the upper limit corresponds to the 75% quantile. The conditional mean is also reported and is equivalent to the conditional median:

```
R> idf <- data.frame(age = gage[i])
R> p <- predict(lm2.9, newdata = idf, level = 0.5,
+                interval = "prediction")
R> colnames(p) <- c("0.5", "0.25", "0.75")
R> p
```

```
   0.5 0.25 0.75
1 46.7 44.5 49.0
2 48.3 46.1 50.5
3 49.8 47.6 52.0
4 51.3 49.1 53.5
```

```
5 52.8 50.6 55.0
6 54.3 52.1 56.5
7 55.9 53.6 58.1
8 57.4 55.2 59.6
```

We now proceed with 80% prediction intervals for constructing the 10% and 90% quantiles, and with 98% prediction intervals corresponding to the 1% and 99% quantiles and repeat the exercise also for the older boys:

```
R> p <- cbind(p, predict(lm2.9, newdata = idf, level = 0.8,
+                        interval = "prediction")[,-1])
R> colnames(p)[4:5] <- c("0.1", "0.9")
R> p <- cbind(p, predict(lm2.9, newdata = idf, level = 0.98,
+                        interval = "prediction")[,-1])
R> colnames(p)[6:7] <- c("0.01", "0.99")
R> p2.9 <- p[, c("0.01", "0.1", "0.25", "0.5",
+                "0.75", "0.9", "0.99")]
R> idf <- data.frame(age = gage[-i])
R> p <- predict(lm9.23, newdata = idf, level = 0.5,
+               interval = "prediction")
R> colnames(p) <- c("0.5", "0.25", "0.75")
R> p <- cbind(p, predict(lm9.23, newdata = idf, level = 0.8,
+                        interval = "prediction")[,-1])
R> colnames(p)[4:5] <- c("0.1", "0.9")
R> p <- cbind(p, predict(lm9.23, newdata = idf, level = 0.98,
+                        interval = "prediction")[,-1])
R> colnames(p)[6:7] <- c("0.01", "0.99")
```

We now reorder the columns of this table and get the following conditional quantiles, estimated under the normal assumption of head circumference:

```
R> p9.23 <- p[, c("0.01", "0.1", "0.25", "0.5",
+                 "0.75", "0.9", "0.99")]
R> round((q2.23 <- rbind(p2.9, p9.23)), 3)
   0.01  0.1 0.25  0.5 0.75  0.9 0.99
1  39.1 42.5 44.5 46.7 49.0 50.9 54.4
2  40.6 44.1 46.1 48.3 50.5 52.5 55.9
3  42.2 45.6 47.6 49.8 52.0 54.0 57.4
4  43.7 47.1 49.1 51.3 53.5 55.5 58.9
5  45.2 48.6 50.6 52.8 55.0 57.0 60.4
6  46.7 50.1 52.1 54.3 56.5 58.5 62.0
7  48.2 51.6 53.6 55.9 58.1 60.1 63.5
8  49.7 53.2 55.2 57.4 59.6 61.6 65.0
1  48.8 50.6 51.7 52.8 54.0 55.1 56.9
2  49.3 51.1 52.1 53.3 54.5 55.5 57.4
3  49.7 51.5 52.6 53.8 55.0 56.0 57.8
4  50.2 52.0 53.1 54.2 55.4 56.5 58.3
5  50.7 52.5 53.5 54.7 55.9 56.9 58.8
6  51.1 53.0 54.0 55.2 56.4 57.4 59.2
7  51.6 53.4 54.5 55.7 56.8 57.9 59.7
8  52.1 53.9 54.9 56.1 57.3 58.4 60.2
9  52.5 54.4 55.4 56.6 57.8 58.8 60.6
10 53.0 54.8 55.9 57.1 58.2 59.3 61.1
11 53.5 55.3 56.4 57.5 58.7 59.8 61.6
12 53.9 55.8 56.8 58.0 59.2 60.2 62.1
13 54.4 56.2 57.3 58.5 59.6 60.7 62.5
```

```
14 54.9 56.7 57.8 58.9 60.1 61.2 63.0
15 55.3 57.2 58.2 59.4 60.6 61.6 63.5
```

We can now superimpose these conditional quantiles on our scatterplot. To do this, we need to write our own little panel function that produces the scatterplot using the `panel.xyplot` function and then adds the just computed conditional quantiles by means of the `panel.lines` function called for every column of `q2.23`.

Figure 12.2 shows parallel lines owing to the fact that the linear model assumes an error variance independent from age; this is the so-called variance homogeneity. Compared to a plot with only a single (mean) regression line, we plotted a whole bunch of conditional distributions here, one for each value of age. Of course, we did so under extremely simplifying assumptions like linearity and variance homogeneity that we're going to drop now.

For the production of a nonparametric version of our growth curves, we start with fitting not only one but multiple quantile regression models, one for each value of τ. We start with the younger boys

```
R> (rq2.9 <- rq(head ~ age, data = db, tau = tau,
+               subset = age < 9))
```

```
Call:
rq(formula = head ~ age, tau = tau, data = db, subset = age <
    9)

Coefficients:
            tau= 0.01 tau= 0.10 tau= 0.25 tau= 0.50 tau= 0.75
(Intercept)     35.63     38.37     41.20     45.01     46.70
age              1.68      1.68      1.55      1.28      1.28
            tau= 0.90 tau= 0.99
(Intercept)     47.69     49.43
age              1.35      1.42

Degrees of freedom: 3193 total; 3191 residual
```

and continue with the older boys

```
R> (rq9.23 <- rq(head ~ age, data = db, tau = tau,
+                subset = age > 9))
```

```
Call:
rq(formula = head ~ age, tau = tau, data = db, subset = age >
    9)

Coefficients:
            tau= 0.01 tau= 0.10 tau= 0.25 tau= 0.50 tau= 0.75
(Intercept)    44.335    46.438     47.60    48.579    49.672
age             0.481     0.469      0.46     0.472     0.477
            tau= 0.90 tau= 0.99
(Intercept)    50.716    52.667
age             0.475     0.465

Degrees of freedom: 3842 total; 3840 residual
```

Naturally, the intercept parameters vary but there is also a considerable variation in the slopes, with the largest value for the 1% quantile regression model for younger boys. The parameters β_τ have to be interpreted with care. In general, they cannot be interpreted on an individual-specific level. A boy who happens to be at the $\tau \times 100\%$ quantile of head circumference conditional

```
R> pfun <- function(x, y, ...) {
+      panel.xyplot(x = x, y = y, ...)
+      if (max(x) <= 9) {
+          apply(q2.23, 2, function(x)
+                panel.lines(gage[i], x[i]))
+      } else {
+          apply(q2.23, 2, function(x)
+                panel.lines(gage[-i], x[-i]))
+      }
+      panel.text(rep(max(db$age), length(tau)),
+                 q2.23[nrow(q2.23),], label = tau, cex = 0.9)
+      panel.text(rep(min(db$age), length(tau)),
+                 q2.23[1,], label = tau, cex = 0.9)
+  }
R> xyplot(head ~ age | cut, data = db, xlab = "Age (years)",
+         ylab = "Head circumference (cm)", pch = 19,
+         scales = list(x = list(relation = "free")),
+         layout = c(2, 1), col = rgb(.1, .1, .1, .1),
+         panel = pfun)
```

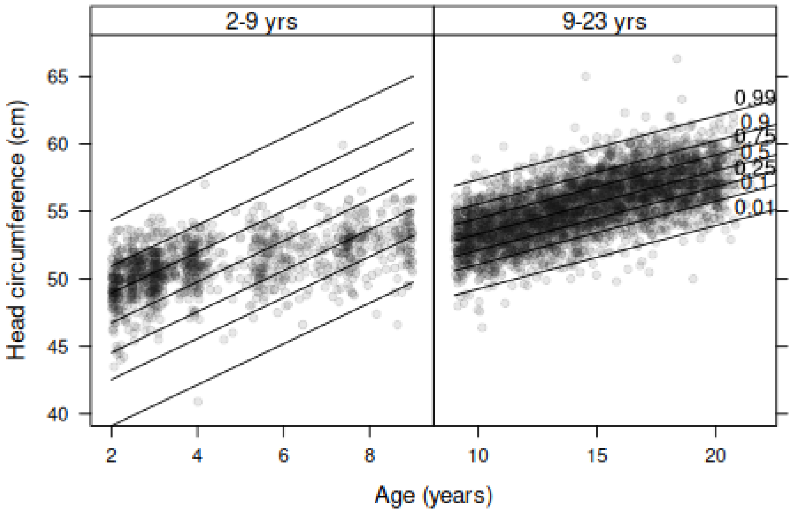

Figure 12.2 Scatterplot of age and head circumference for 5101 Dutch boys with superimposed normal quantiles.

on his age would not be at the same quantile anymore when he gets older. When knowing β_τ, the only conclusion that can be drawn is how the $\tau \times 100\%$ quantile of a population with a specific age differs from the $\tau \times 100\%$ quantile of a population with a different age.

Because the linear functions estimated by linear quantile regression, here in model rq9.23, directly correspond to the conditional quantiles of interest, we can use the predict function to compute the estimated conditional quantiles:

```
R> p2.23 <- rbind(predict(rq2.9,
+                     newdata = data.frame(age = gage[i])),
+                 predict(rq9.23,
+                     newdata = data.frame(age = gage[-i])))
```

It is important to note that these numbers were obtained without assuming anything about the continuous distribution of head circumference given any age. Again, we produce a scatterplot with superimposed quantiles, this time each line corresponds to a specific model. For the sake of comparison with the linear model, we add the linear model quantiles as dashed lines to Figure 12.3. For the older boys, there seems to be almost no difference but the more extreme 1% and 99% quantiles for the younger boys differ considerably. So, at least for the younger boys, we might want to allow for age-specific variability in the distribution of head circumference.

Still, with the quantile regression models shown in Figure 12.3 we assume that the quantiles of head circumference depend on age in a linear way. Additive quantile regression is one way to approach the estimation of non-linear quantile functions. By considering two different models for younger and older boys, we allowed for a certain type of non-linear function in the results shown so far. Additive quantile regression should be able to deal with this problem and we therefore fit these models to all boys simultaneously. For our different choices of τ, we fit one additive quantile regression model using the rqss function from the **quantreg** and allow smooth quantile functions of age via the qss function in the right-hand side of the model formula. Note that we transformed age by the third root prior to model fitting. This does not affect the model since it is a monotone transformation, however, it helps to avoid fitting a function with large derivatives for very young boys resulting in a low penalty parameter λ:

```
R> rqssmod <- vector(mode = "list", length = length(tau))
R> db$lage <- with(db, age^(1/3))
R> for (i in 1:length(tau))
+      rqssmod[[i]] <- rqss(head ~ qss(lage, lambda = 1),
+                          data = db, tau = tau[i])
```

For the analysis of the head circumference, we choose a penalty parameter $\lambda = 1$, which is the default for the qss function. Simply using the default without a careful hyperparameter tuning, for example using crossvalidation or similar procedures, is almost always a mistake. By visual inspection (Figure 12.4) we find this choice appropriate but ask the readers to make a second guess (Exercise 3).

```
R> pfun <- function(x, y, ...) {
+      panel.xyplot(x = x, y = y, ...)
+      if (max(x) <= 9) {
+          apply(q2.23, 2, function(x)
+                panel.lines(gage[i], x[i], lty = 2))
+          apply(p2.23, 2, function(x)
+                panel.lines(gage[i], x[i]))
+      } else {
+          apply(q2.23, 2, function(x)
+                panel.lines(gage[-i], x[-i], lty = 2))
+          apply(p2.23, 2, function(x)
+                panel.lines(gage[-i], x[-i]))
+      }
+      panel.text(rep(max(db$age), length(tau)),
+                 p2.23[nrow(p2.23),], label = tau, cex = 0.9)
+      panel.text(rep(min(db$age), length(tau)),
+                 p2.23[1,], label = tau, cex = 0.9)
+  }
R> xyplot(head ˜ age | cut, data = db, xlab = "Age (years)",
+         ylab = "Head circumference (cm)", pch = 19,
+         scales = list(x = list(relation = "free")),
+         layout = c(2, 1), col = rgb(.1, .1, .1, .1),
+         panel = pfun)
```

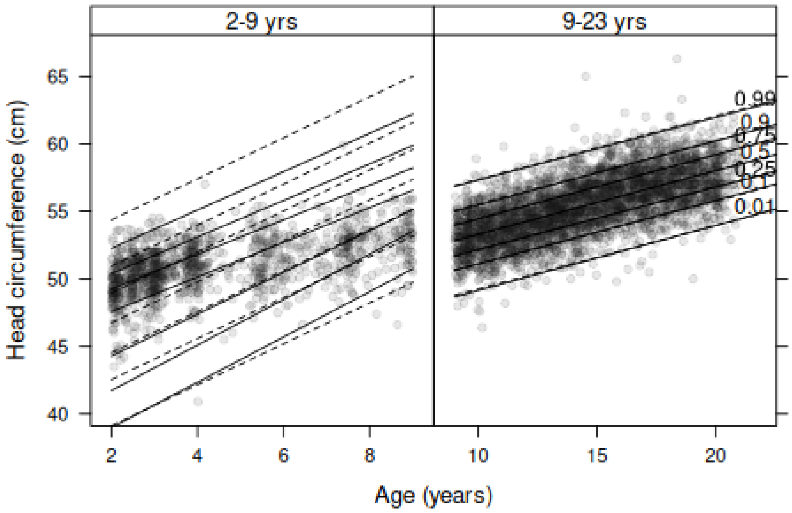

Figure 12.3 Scatterplot of age and head circumference for 5101 Dutch boys with superimposed regression quantiles (solid lines) and normal quantiles (dashed lines).

For a finer grid of age values, we compute the conditional quantiles from the `predict` function:

```
R> gage <- seq(from = min(db$age), to = max(db$age),
+             length = 50)
R> p <- sapply(1:length(tau), function(i) {
+      predict(rqssmod[[i]],
+          newdata = data.frame(lage = gage^(1/3)))
+  })
```

Using very similar code as for plotting linear quantiles, we produce again a scatterplot of age and head circumference but this time overlaid with non-linear regression quantiles. Given that the results from the linear models presented in Figure 12.3 looked pretty convincing, the quantile curves in Figure 12.4 shed a surprising new light on the data. For the younger boys, we expected to see a larger variability than for boys between two and three years old, but in fact the distribution seems to be more complex. The distribution seems to be positively skewed with a heavy lower tail and the degree of skewness varies with age (note that the median is almost linear for boys older than four years).

Also in the right part of Figure 12.4, we see an age-varying skewness, although less pronounced as for the younger boys. The median increases up to 16 years but then the growth rate is much smaller. This does not seem to be the case for the $1\%, 10\%, 90\%$, and 99% quantiles. Note that the discontinuity in the quantiles between the two age groups is only due to the overlapping abscissae.

However, the deviations between the growth curves obtained from a linear model under normality assumption on the one hand and quantile regression on the other hand as shown in Figures 12.3 and 12.4 are hardly dramatic for the head circumference data.

12.4 Summary of Findings

We can conclude that the whole distribution of head circumference changes with age and that assumptions like symmetry and variance homogeneity might be questionable for such type of analysis.

One alternative to the estimation of conditional quantiles is the estimation of conditional distributions. One very interesting parametric approach are generalized additive models for location, scale, and shape (GAMLSS, Rigby and Stasinopoulos, 2005). In Stasinopoulos and Rigby (2007), an analysis of the age and head circumference by means of the **gamlss** package can be found.

One practical problem associated with contemporary methods in quantile regression is quantile crossing. Because we fitted one quantile regression model for each of the quantiles of interest, we cannot guarantee that the conditional quantile functions are monotone, so the 90% quantile may well be larger than the 95% quantile in some cases. Postprocessing of the estimated quantile curves may help in this situation (Dette and Volgushev, 2008).

```
R> pfun <- function(x, y, ...) {
+      panel.xyplot(x = x, y = y, ...)
+      apply(p, 2, function(x) panel.lines(gage, x))
+      panel.text(rep(max(db$age), length(tau)),
+                 p[nrow(p),], label = tau, cex = 0.9)
+      panel.text(rep(min(db$age), length(tau)),
+                 p[1,], label = tau, cex = 0.9)
+  }
R> xyplot(head ~ age | cut, data = db, xlab = "Age (years)",
+         ylab = "Head circumference (cm)", pch = 19,
+         scales = list(x = list(relation = "free")),
+         layout = c(2, 1), col = rgb(.1, .1, .1, .1),
+         panel = pfun)
```

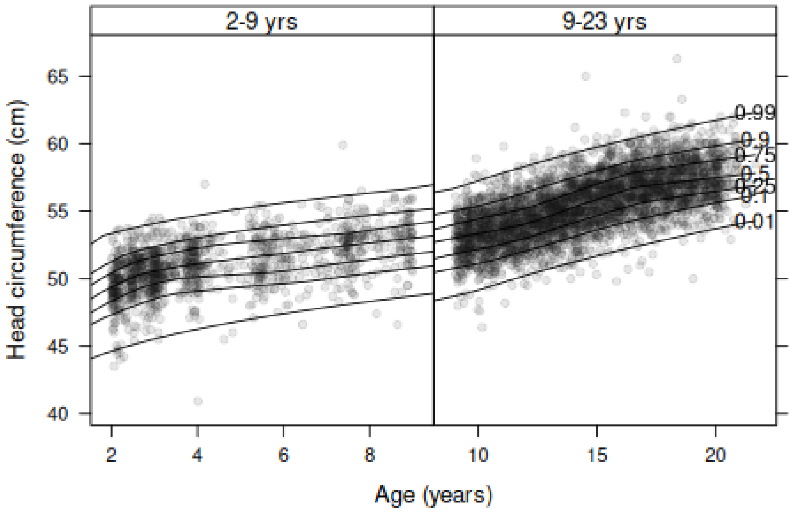

Figure 12.4 Scatterplot of age and head circumference for 5101 Dutch boys with superimposed non-linear regression quantiles.

12.5 Final Comments

When estimating regression models, we have to be aware of the implications of model assumptions when interpreting the results. Symmetry, linearity, and variance homogeneity are among the strongest but common assumptions. Quantile regression, both in its linear and additive formulation, is an intellectually stimulating and practically very useful framework where such assumptions can be relaxed. At a more basic level, one should always ask *Am I really interested in the mean?* before using the regression models discussed in other chapters of this book.

Bibliography

Dette, H. and Volgushev, S. (2008), "Non-crossing non-parametric estimates of quantile curves," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 609–627.

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

Rigby, R. A. and Stasinopoulos, D. M. (2005), "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507–554.

Sarkar, D. (2014), ***lattice**: Lattice Graphics*, URL `http://CRAN.R-project.org/package=lattice`, R package version 0.20-27.

Stasinopoulos, D. M. and Rigby, R. A. (2007), "Generalized additive models for location scale and shape (GAMLSS) in R," *Journal of Statistical Software*, 23, 1–46, URL `http://www.jstatsoft.org/v23/i07`.