Quantile Regression - Chapter 12 on Handbook

Semhar Michael and cps

Quantile Regression

```r
library(gamlss.data)
library(lattice)
library("quantreg")
```

Introduction

During ultrasound examination of an as-yet-unborn baby, anthropometric measurements are taken.

- For a given gestational age one can directly compare, say the femur length of the examined fetus with the femur length of all fetuses in the reference population.
- Too small or too large values may indicate developmental problems and require an intervention.
- From a statistical point of view : what does *too small* or *too* large mean?

Head Circumference for Age - WHO

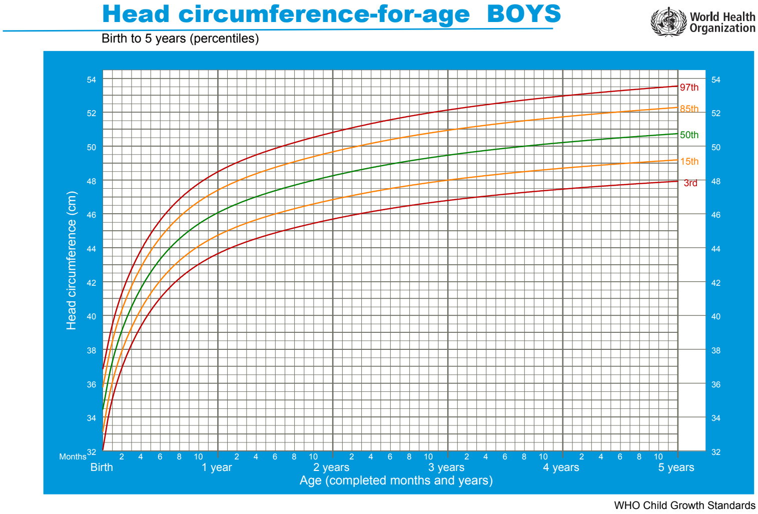


Figure 1: Boys Head Circumference Quantiles

Head Circumference for Age

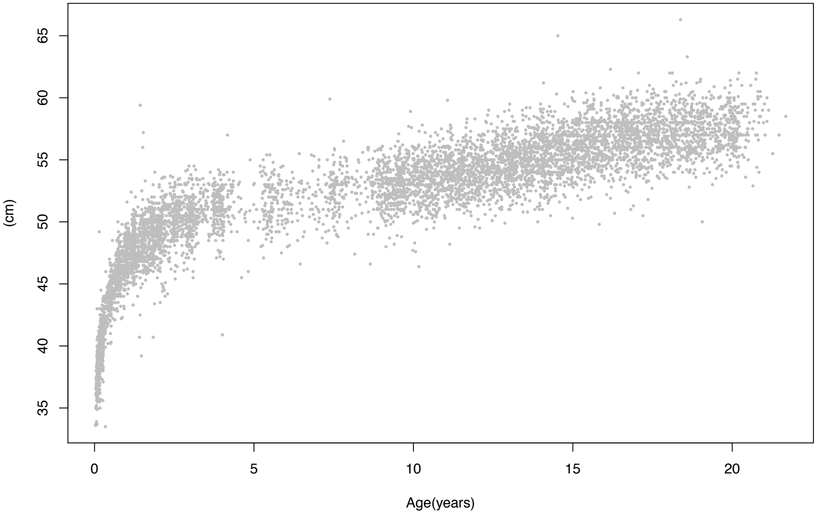
- The data contains head circumference for boys older than 24 months
- Aim: to construct a *growth chart*
- *i.e* Conditional distribution of head circumference given age.
- Age specific quantiles tells us how many boys in the reference population have a smaller head circumference compared to the single boy a physician is looking at.
- Quantile regression - method to estimate conditional quantiles

Head Circumference for Age

```r
#library(gamlss.data)
data(db)
head(db)
dim(db)
plot(db$head ~ db$age, xlab = "Age(months)",
     ylab = "Head circumference",
     pch = 16, cex = 0.5, col = "gray")
```

Head Circumference for Age

```
##   head  age
## 1 33.6 0.03
## 2 33.6 0.04
## 3 33.7 0.04
```



Common regerssion models

- To date, our *Linear* or *additive models* have focused on describing the conditional **mean**, $E(y|x_1, x_2, \ldots, x_q)$ of the response y as a linear or additive function of the explanatory variables $x_1, x_2, \ldots, x_q$.
- For non normal response - link function of the conditional mean is modeled
  - $g(E(y|x_1, x_2, \ldots, x_q))$
- Therefore for a linear model it follows that
  - $y \sim N(\alpha + \beta_1 x_1 + \ldots + \beta_q x_q, \sigma^2)$
  - The conditional $\tau \times 100\%$ quantile for y is $\alpha + \beta_1 x_1 + \ldots + \beta_q x_q + \sigma u_\tau$, where u_τ is the $\tau \times 100\%$ quantile for standard normal
  - For skewed or non-normal distribution the corresponding quantile will be misleading

Linear Quantile Regression

Simple linear quantile regression model (Koenker and Bassett, 1978)

$$y_i = \alpha_\tau + \beta_\tau x_i + \epsilon_{\tau i}$$

where $\epsilon_{\tau i} \sim F_{\tau i}, i = 1, \ldots, n$, subject to $F_{\tau i}(0) = \tau$.

- α_τ and β_τ are the intercept and slope effects and $\tau \in (0, 1)$ is a fixed-known quantile
- $F_{\tau i}$ has no specific distributional assumption except that the distribution function at 0 is τ
- Equivalent to $Q_{y_i}(\tau|x_i) = F_{y_i}^{-1}(\tau|x_i) = \alpha_\tau + \beta_\tau x_i$

Linear Quantile Regression

Minimization problem

- $argmin_{\alpha_\tau \beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - (\alpha_\tau + \beta_\tau x_i))$
  - where $\rho_\tau(z) = z\tau$ for $z \geq 0$ and $z(\tau - 1)$ for $z < 0$

For median $\tau = 0$, $\rho_{0.5}(z) \propto |z|$ therefore

- $argmin_{\alpha_\tau \beta_\tau} \sum_{i=1}^{n} |y_i - (\alpha_\tau + \beta_\tau x_i)|$

The minimization problem above is formulated as a set of linear constraints and estimation of parameters is conducted by linear programming. This will lead to the $\tau \times 100\%$ quantiles of the response variable

- Compare with $argmin_{\alpha\beta} \sum_{i=1}^{n} |y_i - (\alpha + \beta x_i)|^2$ for simple linear regerssion
- Quantile regression is more robust towards extreme outliers as compared to least square regression

Additive Quantile Regression

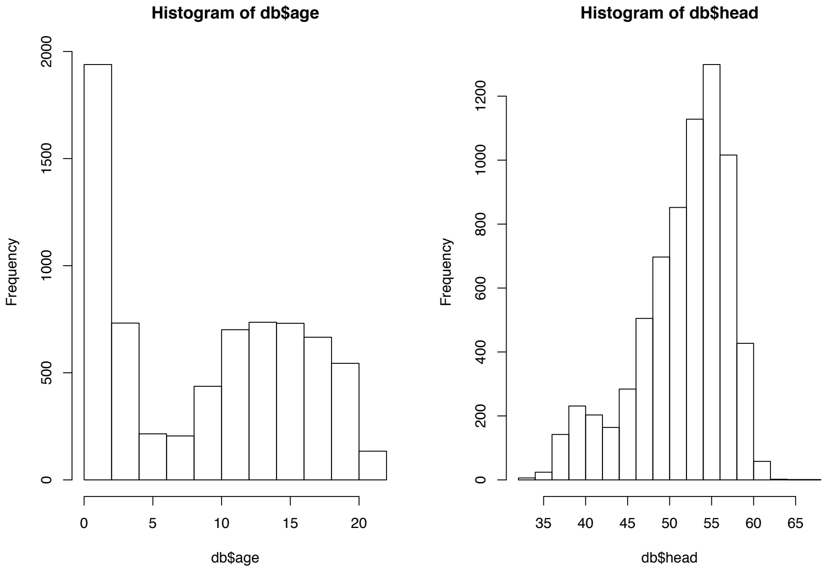
- For cases where non-linear relationship between explanatory variables and quantiles of the response variable
- $Q_{y_i}(\tau|x_i) = f_\tau(x_i)$,
  - where f is a smooth function of x
- The minimization problem is extended by a penality term to

$$argmin_{f_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(x_i)) + \lambda V(f'_\tau),$$

  - where $V(f'_\tau) = \sup \sum |f'_\tau(x_i + 1) - f'_\tau(x_i)|$- the total variation of f'_τ and λ is the tuning parameter
- Solutions are obtained using linear programming (See Koenker et al. 1994 and Koenker 2005)

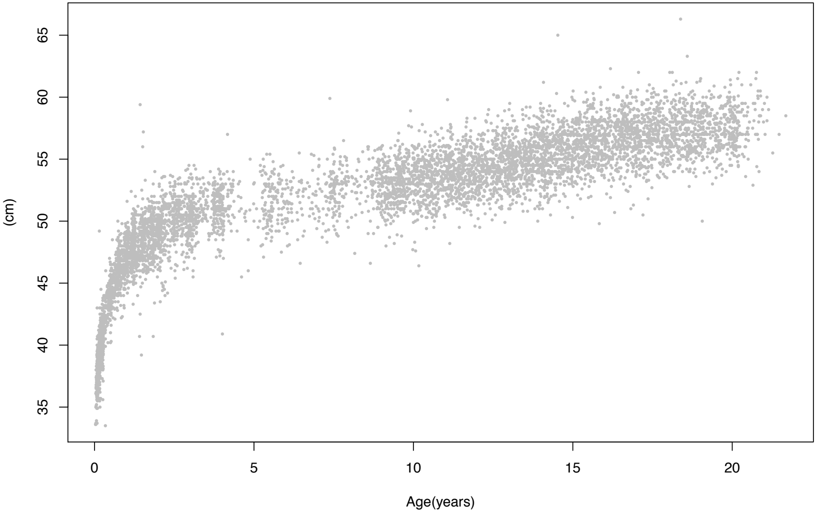
Dutch boys head circumference

```r
layout(matrix(1:2, nrow = 1))
hist(db$age)
hist(db$head)
```



Head Circumference for Age

```
##   head  age
## 1 33.6 0.03
## 2 33.6 0.04
## 3 33.7 0.04
```



Dutch boys head circumference

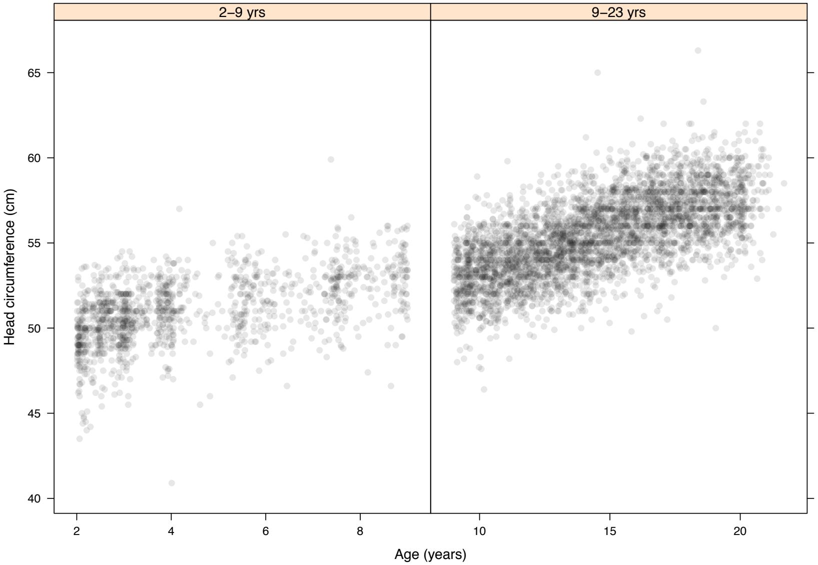
```r
db <- db[db$age>2,] # subset data by age>2
summary(db)
```

```
##       head             age       
##  Min.   :40.90   Min.   : 2.01  
##  1st Qu.:52.30   1st Qu.: 9.08  
##  Median :54.50   Median :12.76  
##  Mean   :54.34   Mean   :12.02  
##  3rd Qu.:56.50   3rd Qu.:16.23  
##  Max.   :66.30   Max.   :21.68  
```

Dutch boys head circumference

```r
#add a cut variable in data to subset data

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

#different scatterplot by age group

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))
```

Dutch boys head circumference



Dutch boys head circumference

Simple linear regression model by age group

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

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

Equivalent to

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

Under the normal assumption the mean is equal to median hence the models can be interpreted as conditional median models under normal assumption

Dutch boys head circumference

Simple linear regression model by age group

```
## (Intercept)         age
##  48.9233698   0.4734876
```

```
## (Intercept)         age
##  48.6194278   0.4689793
```

```
##     I(age < 9)FALSE      I(age < 9)TRUE age:I(age < 9)FALSE
##          48.6201100          48.9233698           0.4689376
##  age:I(age < 9)TRUE
##           0.4734876
```

The model states that within one year, the **average** head circumference for boys less than nine years old increases by 0.473 cm and by 0.469 for older boys.

Dutch boys head circumference - rq

- Relax the distributional assumption (use **rq** function)
- Conditional median ($\tau = 0.5$)

```
rq_med2.9 <- rq(head ~ age, data = db, tau = 0.5,
subset = age < 9)
rq_med2.9$coef
```

```
## (Intercept)         age
##  48.9282511   0.4932735
```

```
rq_med9.23 <- rq(head ~ age, data = db, tau = 0.5,
 subset = age > 9)
rq_med9.23$coef
```

```
## (Intercept)         age
##  48.5791795   0.4717949
```

Dutch boys head circumference - lm vs rq

- Calculate Confidence intervals for the intercept and slope:- Younger boys confidence interval : similar intercept but different slopes using *lm* vs *rq*

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

```
##                             2.5 %   97.5 %
## (Intercept) 48.92337 48.70166 49.14508
```

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

```
##                     2.5 %    97.5 %
## age 0.4734876 0.4282969 0.5186783
```

```
options(warn=-1)# turns off warning message
summary(rq_med2.9, se = "rank")$coef
```

```
##             coefficients   lower bd   upper bd
## (Intercept)   48.9282511 48.7567664 49.1160521
## age            0.4932735  0.4326066  0.5493336
```

```
options(warn=0)# turns on warning message
```

Dutch boys head circumference - lm vs rq

- Calculate Confidence intervals for the intercept and slope:- Older boys confidence interval : similar intercept but different slopes using *lm* vs *rq*

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

```
##                          2.5 %   97.5 %
## (Intercept) 48.61943 48.36341 48.87545
```

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

```
##                    2.5 %    97.5 %
## age 0.4689793 0.4517425 0.4862161
```

```
options(warn=-1)# turns off warning message
summary(rq_med9.23, se = "rank")$coef
```

```
##             coefficients   lower bd   upper bd
## (Intercept)   48.5791795 48.3907933 48.8928025
## age            0.4717949  0.4299378  0.4858946
```

```
options(warn=0)# turns on warning message
```

Dutch boys head circumference - growth curve lm

- Use linear model for construction of growth curves
- Based on the normal linear models, we can compute the quantiles of head circumference for age.
- Here we consider the following values of τ

```
tau <- c(.01, .1, .25, .5, .75, .9, .99)
gage <- c(2:9, 9:23)
i <- 1:8
idf <- data.frame(age = gage[i])
p <- predict(lm2.9, newdata = idf, level = 0.5,
 interval = "prediction") # level - coverage
colnames(p) <- c("0.5", "0.25", "0.75")
p
```

Dutch boys head circumference - growth curve lm

```
##        0.5     0.25     0.75
## 1 49.87034 48.69777 51.04292
## 2 50.34383 49.17165 51.51602
## 3 50.81732 49.64533 51.98931
## 4 51.29081 50.11880 52.46282
## 5 51.76430 50.59206 52.93653
## 6 52.23778 51.06512 53.41044
## 7 52.71127 51.53797 53.88457
## 8 53.18476 52.01062 54.35889
```

Dutch boys head circumference

Find 80% and 98% prediction intervals

```
p <- cbind(p, predict(lm2.9, newdata = idf, level = 0.8,
 interval = "prediction")[,-1])
colnames(p)[4:5] <- c("0.1", "0.9")
p <- cbind(p, predict(lm2.9, newdata = idf, level = 0.98,
 interval = "prediction")[,-1])
colnames(p)[6:7] <- c("0.01", "0.99")
p2.9 <- p[, c("0.01", "0.1", "0.25", "0.5",
 "0.75", "0.9", "0.99")]
head(p2.9)
```

```
##       0.01      0.1     0.25      0.5     0.75      0.9     0.99
## 1 45.82205 47.64188 48.69777 49.87034 51.04292 52.09881 53.91864
## 2 46.29691 48.11612 49.17165 50.34383 51.51602 52.57155 54.39076
## 3 46.77105 48.58997 49.64533 50.81732 51.98931 53.04467 54.86359
## 4 47.24448 49.06342 50.11880 51.29081 52.46282 53.51819 55.33713
## 5 47.71720 49.53649 50.59206 51.76430 52.93653 53.99210 55.81139
## 6 48.18921 50.00916 51.06512 52.23778 53.41044 54.46640 56.28636
```

Dutch boys head circumference

Repeate the same for older boys

```
idf <- data.frame(age = gage[-i])
p <- predict(lm9.23, newdata = idf, level = 0.5,
interval = "prediction")
colnames(p) <- c("0.5", "0.25", "0.75")
p <- cbind(p, predict(lm9.23, newdata = idf, level = 0.8,
 interval = "prediction")[,-1])
colnames(p)[4:5] <- c("0.1", "0.9")
p <- cbind(p, predict(lm9.23, newdata = idf, level = 0.98,
  interval = "prediction")[,-1])
colnames(p)[6:7] <- c("0.01", "0.99")
p9.23 <- p[, c("0.01", "0.1", "0.25", "0.5",
 "0.75", "0.9", "0.99")]
p9.23
```

Dutch boys head circumference

Quantiles for older boys

```
##        0.01      0.1     0.25      0.5     0.75      0.9     0.99
## 1  48.78475 50.60668 51.66479 52.84024 54.01569 55.07381 56.89574
## 2  49.25424 51.07594 52.13392 53.30922 54.48452 55.54250 57.36420
## 3  49.72363 51.54515 52.60302 53.77820 54.95338 56.01125 57.83277
## 4  50.19292 52.01430 53.07209 54.24718 55.42227 56.48006 58.30143
## 5  50.66211 52.48339 53.54113 54.71616 55.89119 56.94893 58.77021
## 6  51.13119 52.95243 54.01014 55.18514 56.36014 57.41785 59.23908
## 7  51.60017 53.42141 54.47912 55.65412 56.82912 57.88683 59.70806
## 8  52.06905 53.89033 54.94807 56.12310 57.29813 58.35586 60.17714
## 9  52.53782 54.35919 55.41699 56.59208 57.76717 58.82496 60.64633
## 10 53.00649 54.82800 55.88588 57.06106 58.23624 59.29411 61.11562
## 11 53.47506 55.29676 56.35473 57.53003 58.70533 59.76331 61.58501
## 12 53.94352 55.76545 56.82356 57.99901 59.17446 60.23258 62.05450
## 13 54.41188 56.23409 57.29236 58.46799 59.64362 60.70190 62.52410
## 14 54.88014 56.70267 57.76113 58.93697 60.11281 61.17127 62.99380
## 15 55.34830 57.17120 58.22988 59.40595 60.58203 61.64071 63.46361
```

Dutch boys head circumference

Conditional quantiles estimated under the normal assumption of head circumference

```
q2.23 <- rbind(p2.9, p9.23)
head(round(q2.23, 3), n = 14)
```

```
##     0.01    0.1   0.25    0.5   0.75    0.9   0.99
## 1 45.822 47.642 48.698 49.870 51.043 52.099 53.919
## 2 46.297 48.116 49.172 50.344 51.516 52.572 54.391
## 3 46.771 48.590 49.645 50.817 51.989 53.045 54.864
## 4 47.244 49.063 50.119 51.291 52.463 53.518 55.337
## 5 47.717 49.536 50.592 51.764 52.937 53.992 55.811
## 6 48.189 50.009 51.065 52.238 53.410 54.466 56.286
## 7 48.661 50.481 51.538 52.711 53.885 54.941 56.762
## 8 49.131 50.953 52.011 53.185 54.359 55.416 57.238
## 1 48.785 50.607 51.665 52.840 54.016 55.074 56.896
## 2 49.254 51.076 52.134 53.309 54.485 55.543 57.364
## 3 49.724 51.545 52.603 53.778 54.953 56.011 57.833
## 4 50.193 52.014 53.072 54.247 55.422 56.480 58.301
## 5 50.662 52.483 53.541 54.716 55.891 56.949 58.770
## 6 51.131 52.952 54.010 55.185 56.360 57.418 59.239
```

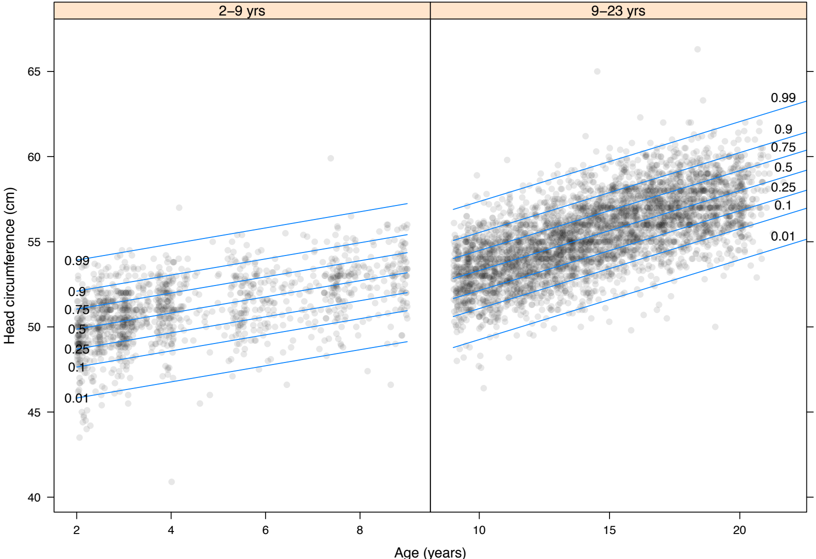
Dutch boys head circumference

Superimpose these conditional quantiles on our scatterplot

```
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)
}
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)
```

Dutch boys head circumference

- Parallel lines owing to the fact that the linear model assumes an error variance independent from age- variance homogeneity



Dutch boys head circumference - growth curves rq

Nonparametric version of our growth curves

```
rq2.9 <- rq(head ~ age, data = db, tau = tau,
 subset = age < 9)
rq2.9$coef
```

```
##               tau= 0.01  tau= 0.10  tau= 0.25  tau= 0.50  tau= 0.75
## (Intercept) 43.2992424 46.9331190 48.0224215 48.9282511 50.1110357
## age          0.6515152  0.4501608  0.4484305  0.4932735  0.4584041
##              tau= 0.90  tau= 0.99
## (Intercept) 50.765014 52.6367698
## age          0.523416  0.4467354
```

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

```
##               tau= 0.01  tau= 0.10  tau= 0.25  tau= 0.50  tau= 0.75
## (Intercept) 44.3351899 46.4375451 47.5965517 48.5791795 49.6719626
## age          0.4810127  0.4693141  0.4597701  0.4717949  0.4766355
##               tau= 0.90  tau= 0.99
## (Intercept) 50.7155801 52.6674762
## age          0.4751381  0.4646251
```

Dutch boys head circumference- growth curves rq

Nonparametric version of our growth curves - prediction

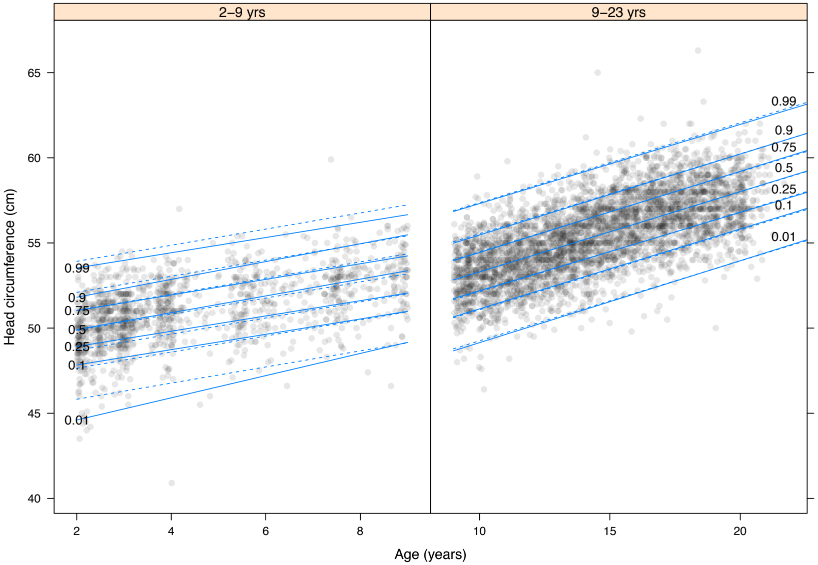
```
p2.23 <- rbind(predict(rq2.9,
  newdata = data.frame(age = gage[i])),
  predict(rq9.23,
  newdata = data.frame(age = gage[-i])))
head(p2.23)
```

```
##   tau= 0.01 tau= 0.10 tau= 0.25 tau= 0.50 tau= 0.75 tau= 0.90 tau= 0.99
## 1  44.60227  47.83344  48.91928  49.91480  51.02784  51.81185  53.53024
## 2  45.25379  48.28360  49.36771  50.40807  51.48625  52.33526  53.97698
## 3  45.90530  48.73376  49.81614  50.90135  51.94465  52.85868  54.42371
## 4  46.55682  49.18392  50.26457  51.39462  52.40306  53.38209  54.87045
## 5  47.20833  49.63408  50.71300  51.88789  52.86146  53.90551  55.31718
## 6  47.85985  50.08424  51.16143  52.38117  53.31986  54.42893  55.76392
```

Dutch boys head circumference - growth curves rq

```
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)
}
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)
```

Dutch boys head circumference - growth curves rq



Dutch boys head circumference - non-linear qr

Non-linear quantile regression (use *rqss* function)

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

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

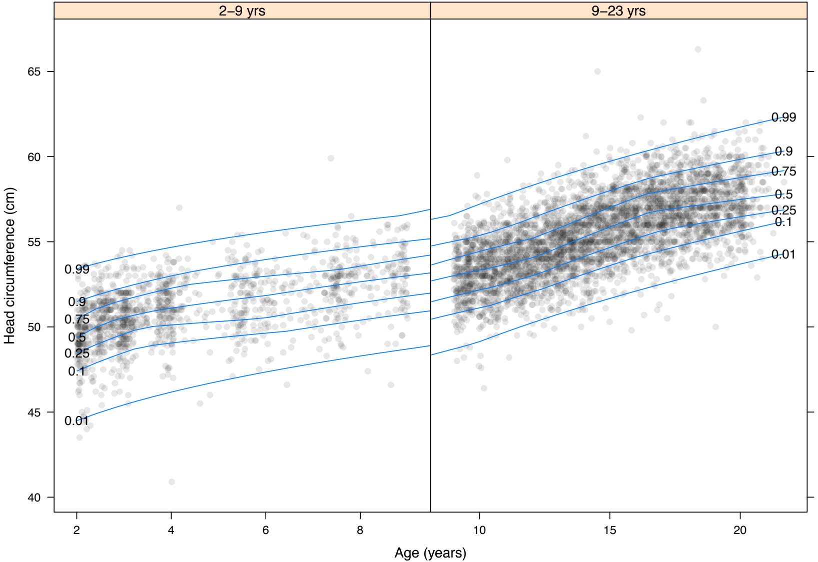
Dutch boys head circumference

Non-linear quantile regression

```
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)
}
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)
```

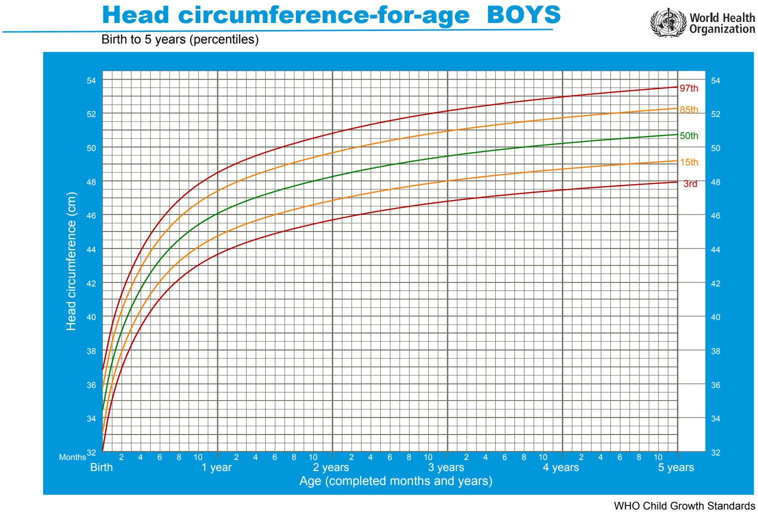
Dutch boys head circumference

Non-linear quantile regression



Playing with the whole db data

Let us now try to replicate the WHO plot



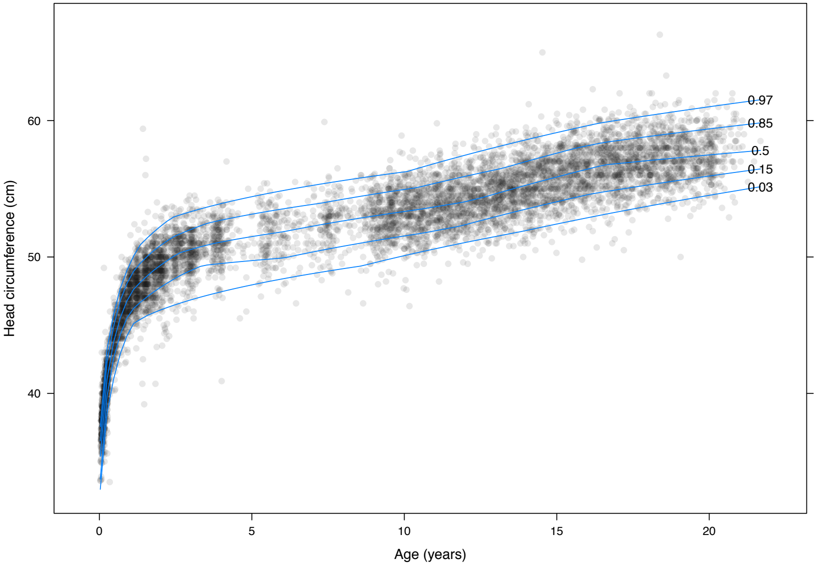
Playing with the whole db data

```
## use the tau values as given in the above plot
library(gamlss.data)
data(db)
db2 <- db
tau <- c(.03, .15, .5, .85, .97)

rqssmod <- vector(mode = "list", length = length(tau))
db2$lage <- with(db2, age^(1/3))
for (i in 1:length(tau))
 rqssmod[[i]] <- rqss(head ~ qss(lage, lambda = 1),
 data = db2, tau = tau[i])

gage <- seq(from = min(db2$age), to = max(db2$age), length = 100)
p <- sapply(1:length(tau), function(i) { predict(rqssmod[[i]],
    newdata = data.frame(lage = gage^(1/3)))
  })
```

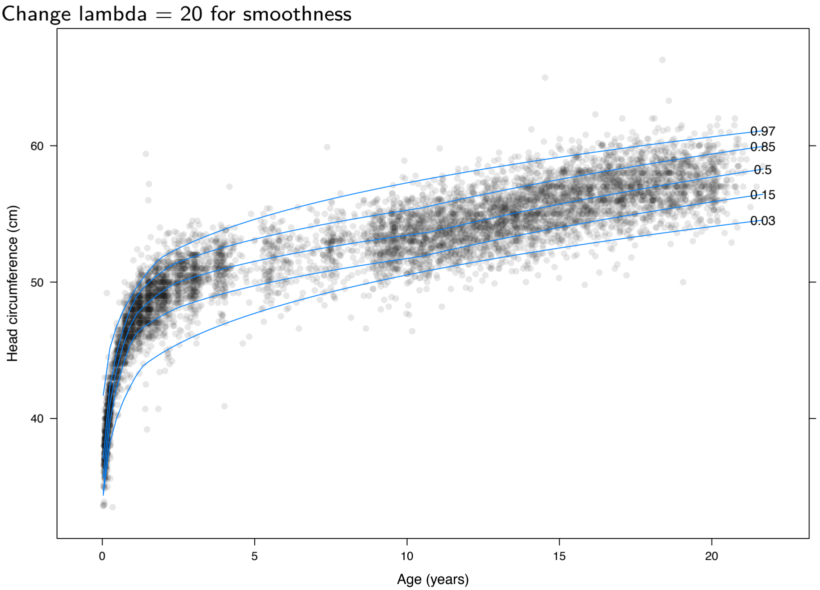
Playing with the whole db data

```
pfun <- function(x, y, ...) {
 panel.xyplot(x = x, y = y, ...)
 apply(p, 2, function(x) panel.lines(gage, x))
 panel.text(rep(max(db2$age), length(tau)),
 p[nrow(p),], label = tau, cex = 0.9)
 #panel.text(rep(min(db2$age), length(tau)),
 #p[1,], label = tau, cex = 0.9)
}
xyplot(head ~ age, data = db2, xlab = "Age (years)",
 ylab = "Head circumference (cm)", pch = 19,
 scales = list(x = list(relation = "free")),
 layout = c(1, 1), col = rgb(.1, .1, .1, .1),
 panel = pfun)
```

Playing with the whole db data



Playing with the whole db data



Quantile regression final remark

- 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.