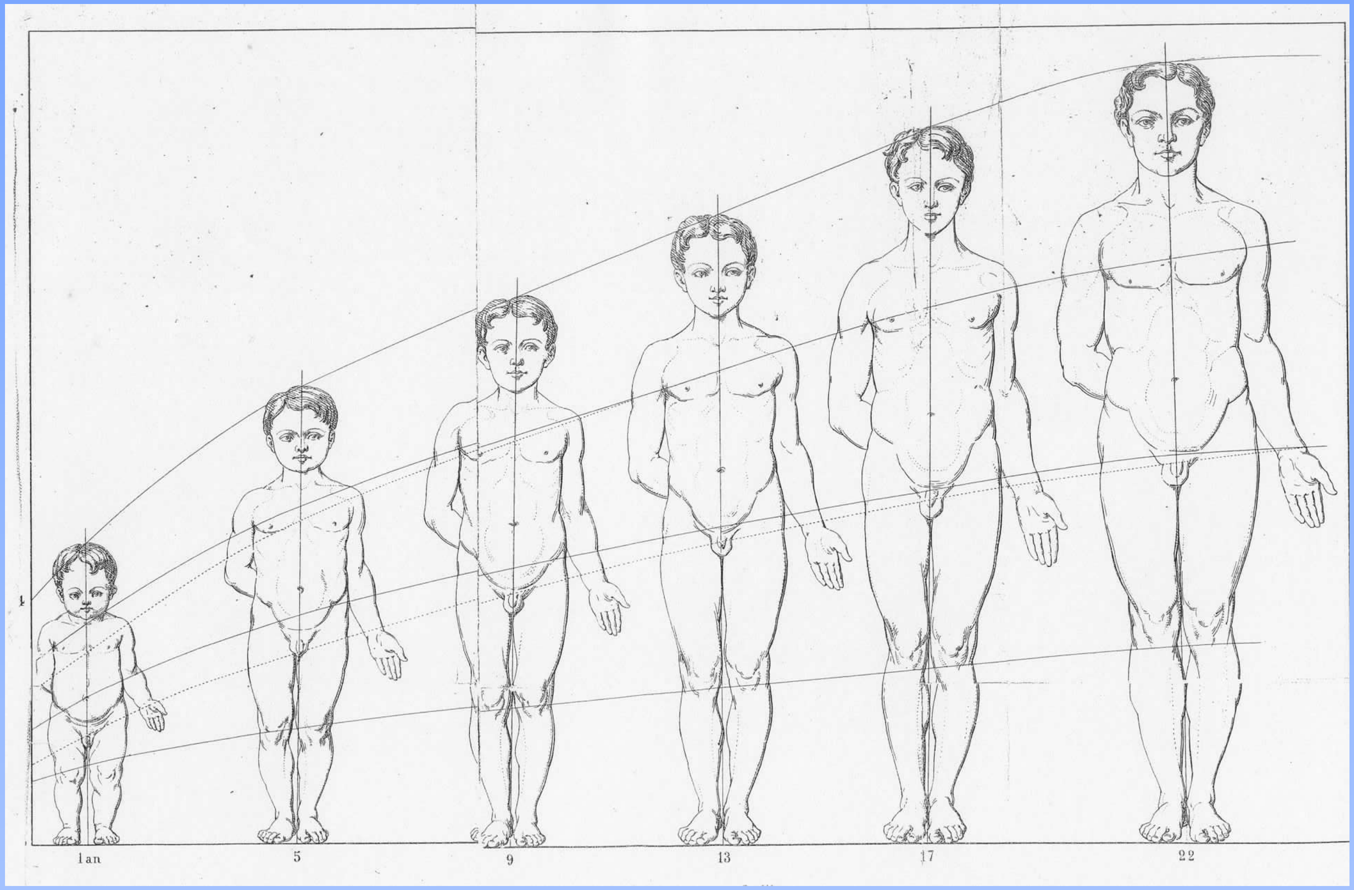
Quantile Regression Methods for Reference Growth Charts

Roger Koenker
University of Illinois at Urbana-Champaign

ASA Workshop on Nonparametric Statistics
Texas A&M, January 15, 2005

Based on joint work with:
Ying Wei, Columbia Biostatistics
Anneli Pere, Children's Hospital, Helsinki
Xuming He, UIUC.

Quetelet's (1871) Growth Chart



Penalized Maximum Likelihood Estimation

- References: Cole (1988), Cole and Green (1992), and Carey(2002)
- Data: $\{Y_i(t_{i,j}) : \; j = 1, \dots, J_i, \; i = 1, \dots, n.\}$
- Model: $Z(t) = \frac{(Y(t)/\mu(t))^{\lambda(t)} - 1}{\lambda(t)\sigma(t)} \sim \mathcal{N}(0,1)$
- Estimation:

$$\max \ell(\lambda, \mu, \sigma) - \nu_\lambda \int (\lambda''(t))^2 dt - \nu_\mu \int (\mu''(t))^2 dt - \nu_\sigma \int (\sigma''(t))^2 dt,$$

$$\ell(\lambda, \mu, \sigma) = \sum_{i=1}^{n} [\lambda(t_i) \log(Y(t_i)/\mu(t_i)) - \log \sigma(t_i) - \tfrac{1}{2} Z^2(t_i)],$$

Quantiles as Argmins

The τth quantile of a random variable Y having distribution function F is:

$$\alpha(\tau) = \text{argmin} \int \rho_\tau(y - \alpha) dF(y)$$

where

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)).$$

Quantiles as Argmins

The τth quantile of a random variable Y having distribution function F is:

$$\alpha(\tau) = \text{argmin} \int \rho_\tau(y - \alpha) dF(y)$$

where

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)).$$

The τth *sample* quantile is thus:

$$\begin{aligned} \hat{\alpha}(\tau) &= \text{argmin} \int \rho_\tau(y - \alpha) dF_n(y) \\ &= \text{argmin}\, n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha) \end{aligned}$$

Quantile Regression

The τth conditional quantile function of $Y|X = x$ is

$$g(\tau|x) = \text{argmin}_{g \in \mathcal{G}} \int \rho_\tau(y - g(x))dF$$

A natural estimator of $g(\tau|x)$ is

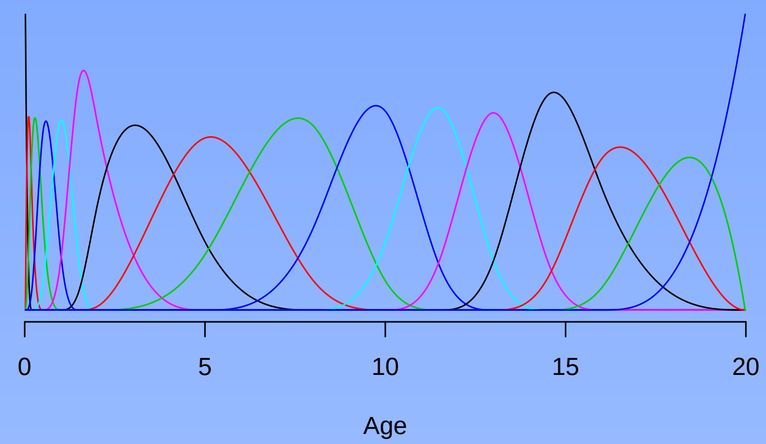
$$\hat{g}(\tau|x) = \text{argmin}_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i))$$

with $\mathcal{G}$ chosen as a finite dimensional linear space,

$$g(x) = \sum_{j=1}^{p} \varphi_j(x)\beta_j.$$

Choice of Basis

There are many possible choices for the basis expansion $\{\varphi_j\}$. We opt for the (very conventional) cubic B-spline functions:



In R quantile regression models can be estimated with the command `fit <- rq(y` $\sim$ `bs(x,knots=knots),tau = 1:9/10)` Similar functionality in SAS is coming "real soon now."

Data

- Longitudinal measurements on height for 2514 Finnish children,

Data

- Longitudinal measurements on height for 2514 Finnish children,
- 1143 boys, 1162 girls – all healthy, full-term, singleton births,

Data

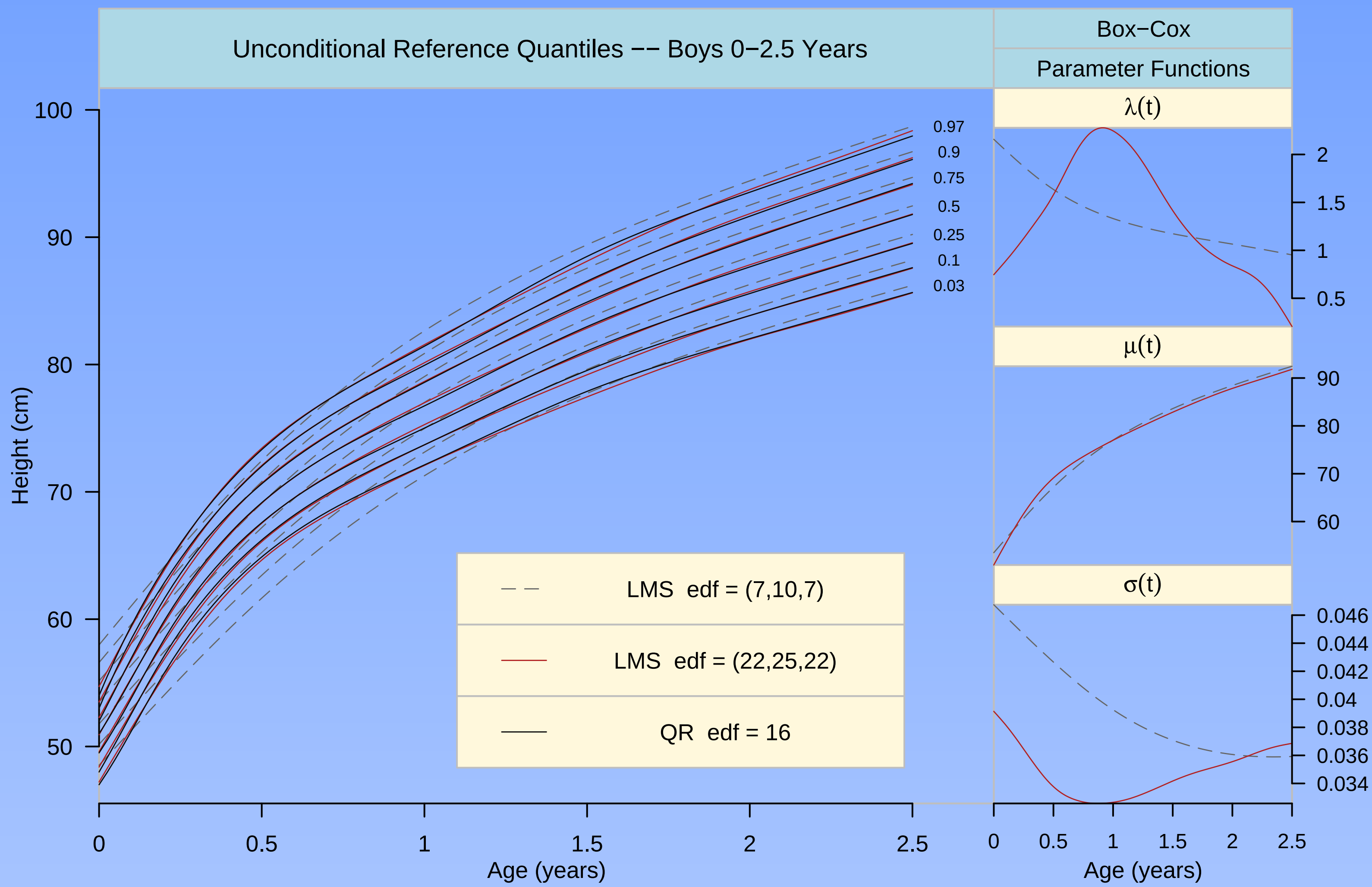
- Longitudinal measurements on height for 2514 Finnish children,
- 1143 boys, 1162 girls – all healthy, full-term, singleton births,
- About 20 measurements per child,

Data

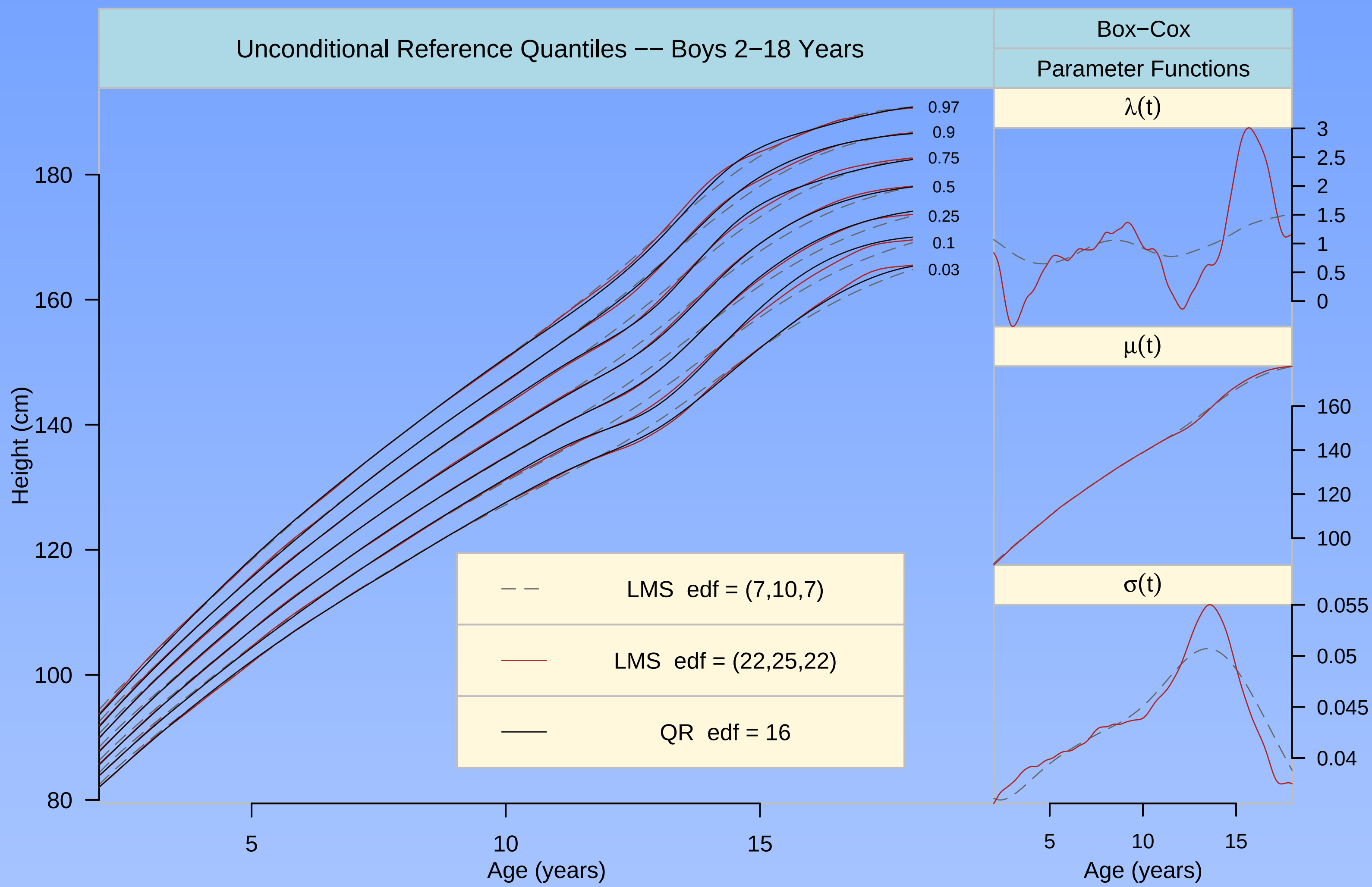
- Longitudinal measurements on height for 2514 Finnish children,
- 1143 boys, 1162 girls – all healthy, full-term, singleton births,
- About 20 measurements per child,
- Two cohorts: 1096 born between 1959-61, 1209 born between 1968-72

Data

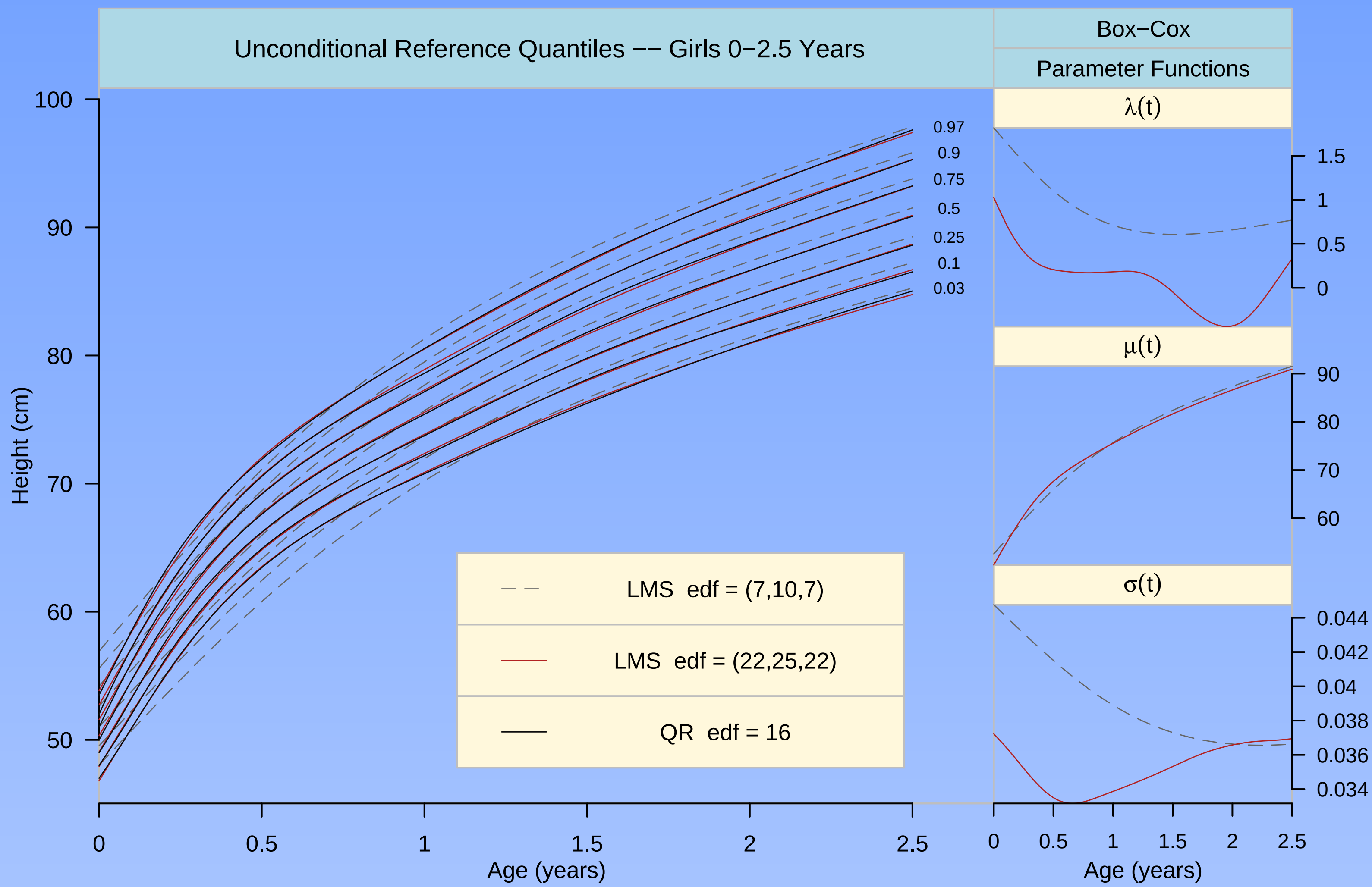
- Longitudinal measurements on height for 2514 Finnish children,
- 1143 boys, 1162 girls – all healthy, full-term, singleton births,
- About 20 measurements per child,
- Two cohorts: 1096 born between 1959-61, 1209 born between 1968-72
- Sample constitutes 0.5 percent of Finns born in these periods.

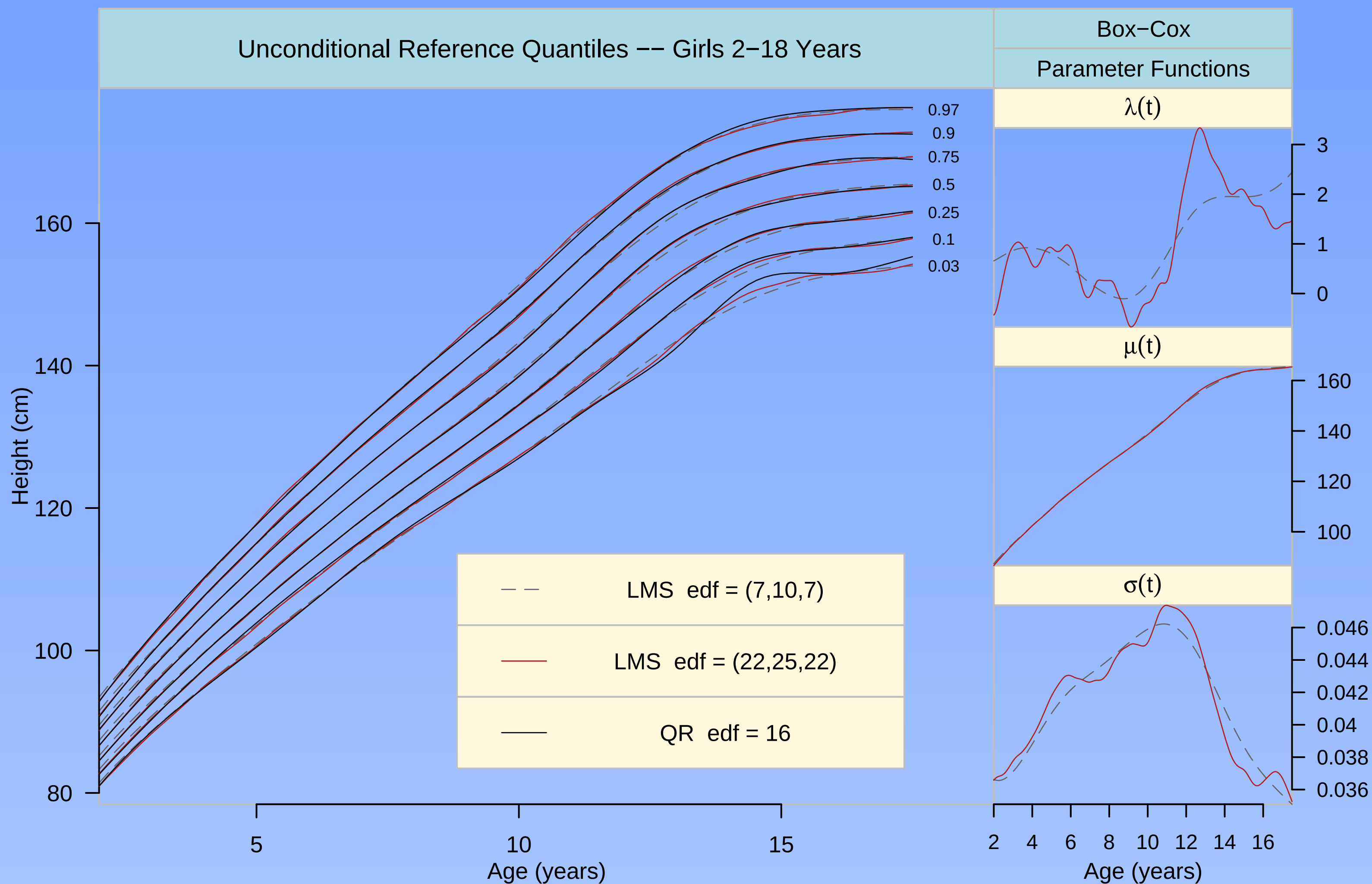


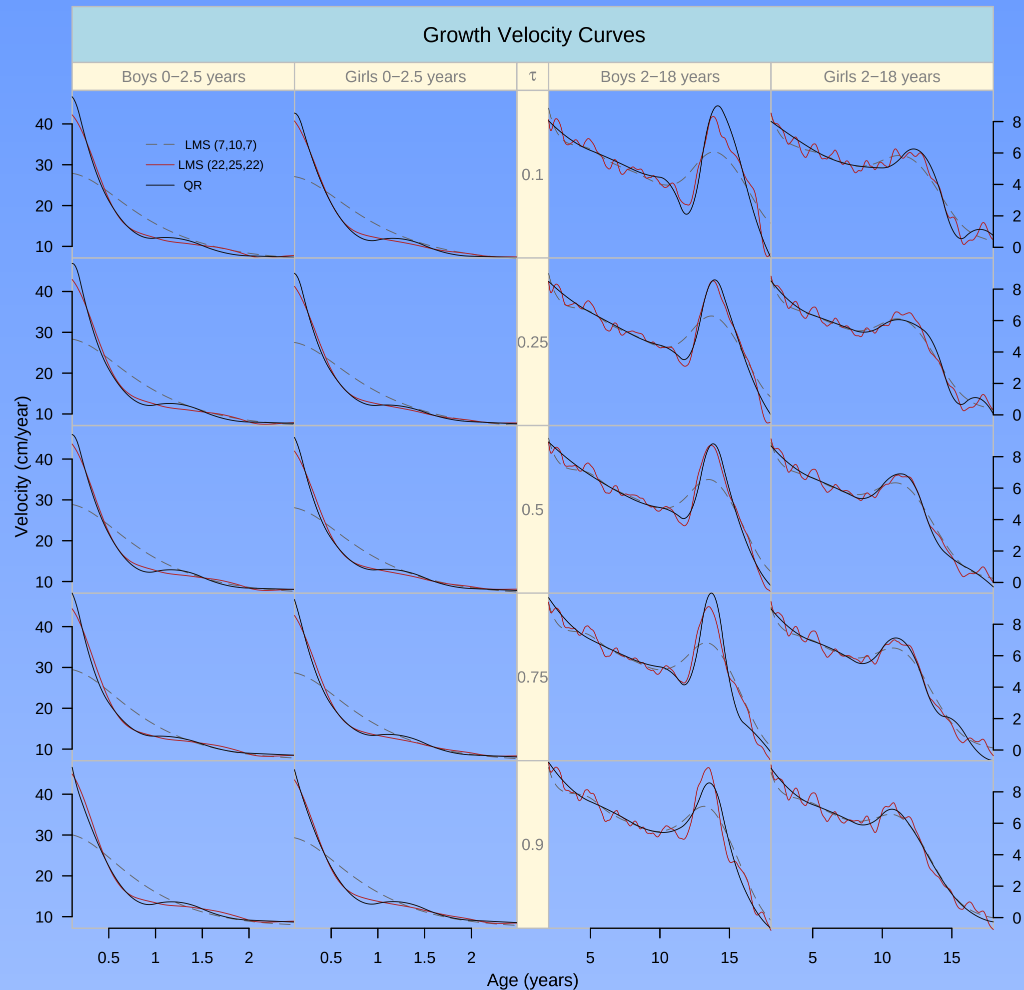
Unconditional Reference Quantiles −− Boys 0−2.5 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
100
90
80
70
60
50
0.97
0.9
0.75
0.5
0.25
0.1
0.03
LMS edf = (7,10,7)
LMS edf = (22,25,22)
QR edf = 16
2
1.5
1
0.5
90
80
70
60
0.046
0.044
0.042
0.04
0.038
0.036
0.034
0
0.5
1
1.5
2
2.5
Age (years)
0
0.5
1
1.5
2
2.5
Age (years)

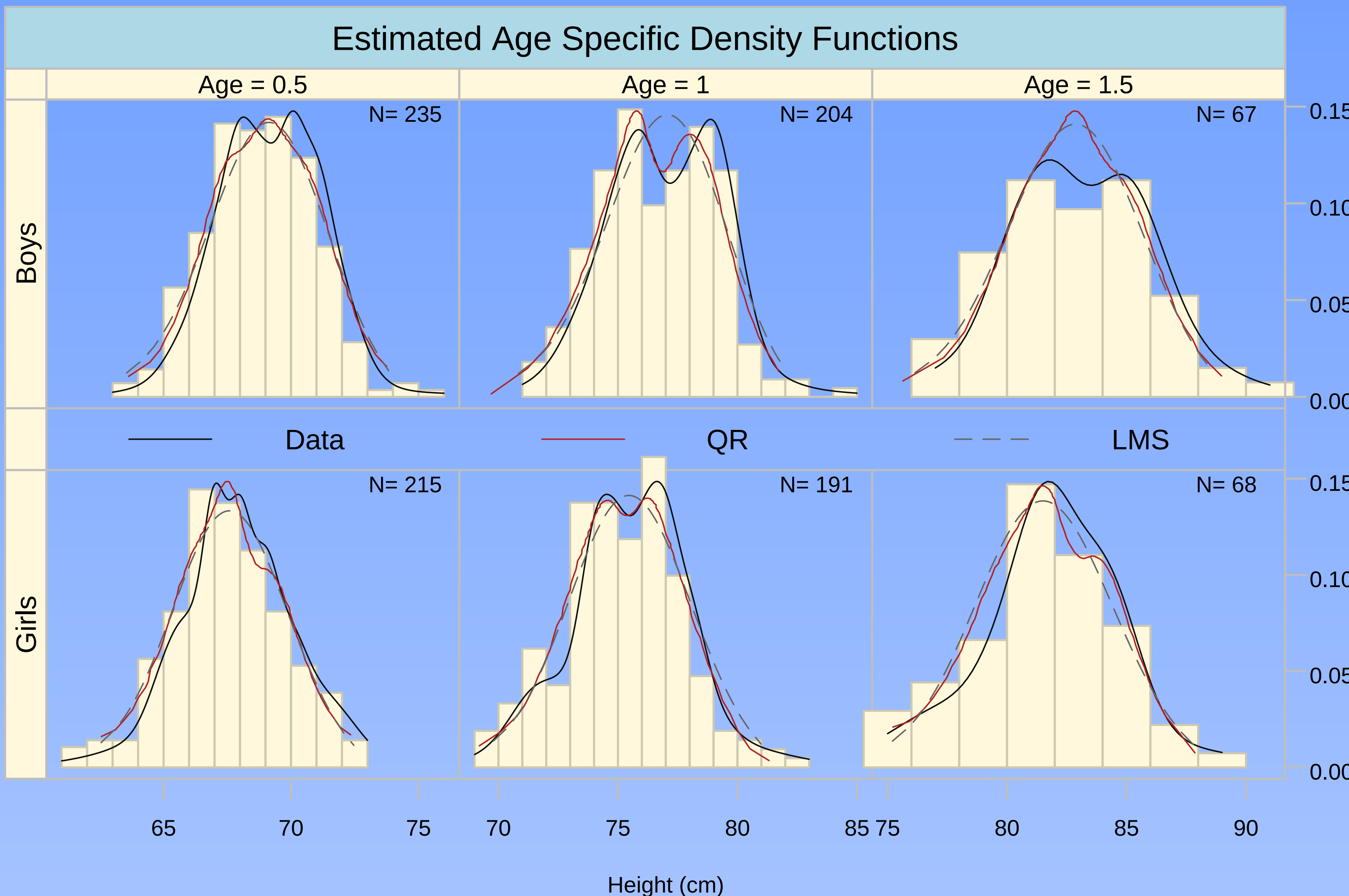


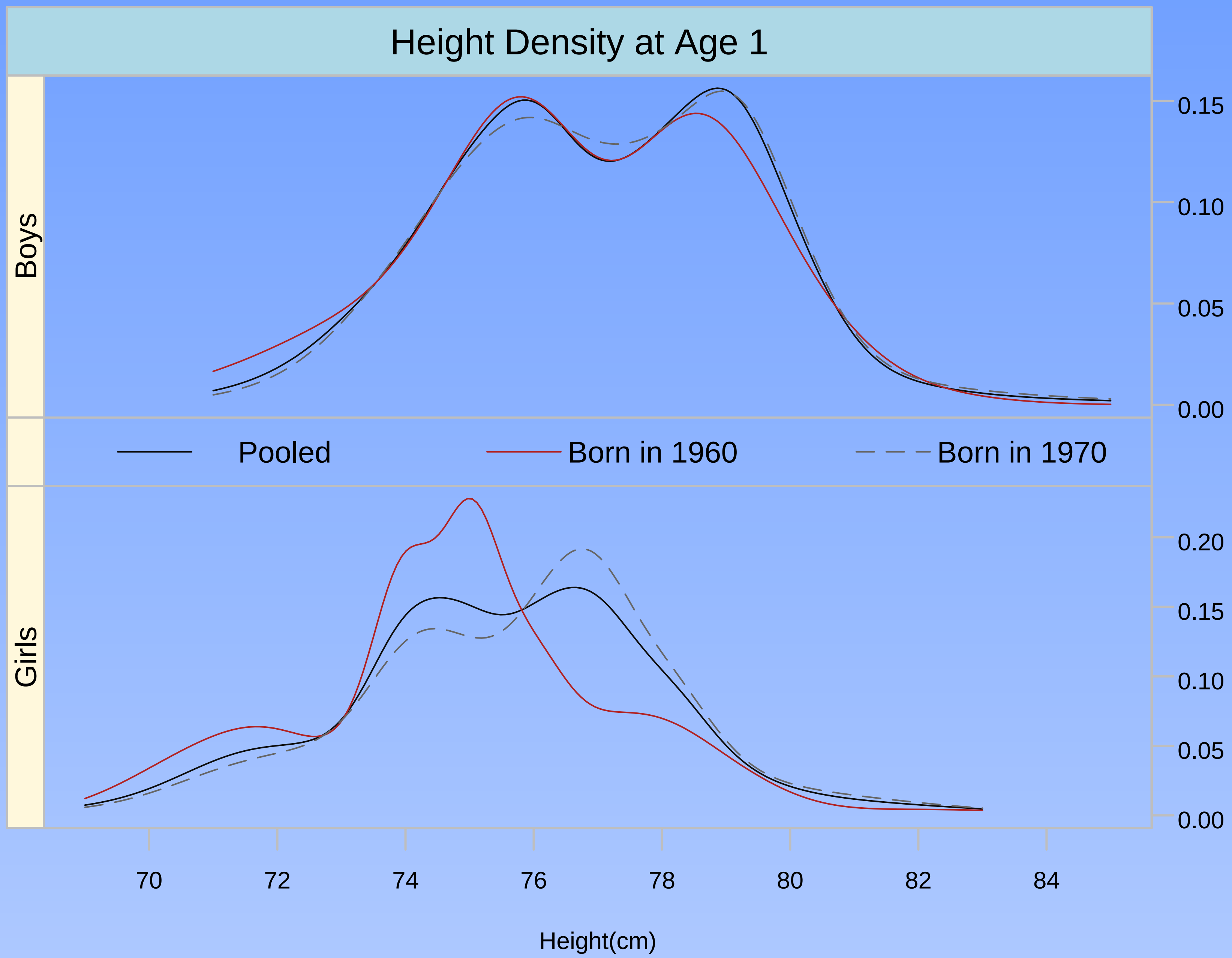
Unconditional Reference Quantiles -- Boys 2-18 Years
Box-Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
180
160
140
120
100
80
5
10
15
Age (years)
0.97
0.9
0.75
0.5
0.25
0.1
0.03
LMS edf = (7,10,7)
LMS edf = (22,25,22)
QR edf = 16
3
2.5
2
1.5
1
0.5
0
160
140
120
100
0.055
0.05
0.045
0.04


Unconditional Reference Quantiles −− Girls 0−2.5 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
Age (years)
0.97
0.9
0.75
0.5
0.25
0.1
0.03
LMS edf = (7,10,7)
LMS edf = (22,25,22)
QR edf = 16
Age (years)


Unconditional Reference Quantiles −− Girls 2−18 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
Age (years)
0.97
0.9
0.75
0.5
0.25
0.1
0.03
LMS edf = (7,10,7)
LMS edf = (22,25,22)
QR edf = 16
Age (years)


Growth Velocity Curves
Boys 0−2.5 years
Girls 0−2.5 years
τ
Boys 2−18 years
Girls 2−18 years
LMS (7,10,7)
LMS (22,25,22)
QR
0.1
0.25
0.5
0.75
0.9
Velocity (cm/year)
Age (years)


Estimated Age Specific Density Functions
Age = 0.5
Age = 1
Age = 1.5
Boys
Girls
N= 235
N= 204
N= 67
N= 215
N= 191
N= 68
Data
QR
LMS
0.00
0.05
0.10
0.15
65
70
75
70
75
80
85
75
80
85
90
Height (cm)


Height Density at Age 1
Boys
Girls
Pooled
Born in 1960
Born in 1970
0.00
0.05
0.10
0.15
0.20
70
72
74
76
78
80
82
84
Height(cm)

Conditioning on Prior Growth

It is often important to condition not only on age, but also on prior growth and possibly on other covariates. Autoregressive models are natural, but complicated due to the irregular spacing of typical longitudinal measurements.

- Data: $\{Y_i(t_{i,j}) : \; j = 1, \ldots, J_i, \; i = 1, \ldots, n.\}$
- Model:

$$\begin{aligned} Q_{Y_i(t_{i,j})}(\tau \quad | \quad & t_{i,j}, Y_i(t_{i,j-1}), x_i) = g_\tau(t_{i,j}) \\ + \quad & [\alpha(\tau) + \beta(\tau)(t_{i,j} - t_{i,j-1})] Y_i(t_{i,j-1}) + x_i^\top \gamma(\tau). \end{aligned}$$

AR Components of the Infant Conditional Growth Model

τ	**Boys**			**Girls**		
	$\hat{\alpha}(\tau)$	$\hat{\beta}(\tau)$	$\hat{\gamma}(\tau)$	$\hat{\alpha}(\tau)$	$\hat{\beta}(\tau)$	$\hat{\gamma}(\tau)$
0.03	0.845 (0.020)	0.147 (0.011)	0.024 (0.011)	0.809 (0.024)	0.135 (0.011)	0.042 (0.010)
0.1	0.787 (0.020)	0.159 (0.007)	0.036 (0.007)	0.757 (0.022)	0.153 (0.007)	0.054 (0.009)
0.25	0.725 (0.019)	0.170 (0.006)	0.051 (0.009)	0.685 (0.021)	0.163 (0.006)	0.061 (0.008)
0.5	0.635 (0.025)	0.173 (0.009)	0.060 (0.013)	0.612 (0.027)	0.175 (0.008)	0.070 (0.009)
0.75	0.483 (0.029)	0.187 (0.009)	0.063 (0.017)	0.457 (0.027)	0.183 (0.012)	0.094 (0.015)
0.9	0.422 (0.024)	0.213 (0.016)	0.070 (0.017)	0.411 (0.030)	0.201 (0.015)	0.100 (0.018)
0.97	0.383 (0.024)	0.214 (0.016)	0.077 (0.018)	0.400 (0.038)	0.232 (0.024)	0.086 (0.027)

AR Components of the Childrens' Conditional Growth Model

τ	Boys			Girls		
	$\hat{\alpha}(\tau)$	$\hat{\beta}(\tau)$	$\hat{\gamma}(\tau)$	$\hat{\alpha}(\tau)$	$\hat{\beta}(\tau)$	$\hat{\gamma}(\tau)$
0.03	0.976 (0.010)	0.036 (0.002)	0.011 (0.013)	0.993 (0.012)	0.033 (0.002)	0.006 (0.015)
0.1	0.980 (0.005)	0.039 (0.001)	0.022 (0.007)	0.989 (0.006)	0.039 (0.001)	0.008 (0.007)
0.25	0.978 (0.006)	0.042 (0.001)	0.021 (0.006)	0.986 (0.005)	0.042 (0.001)	0.019 (0.006)
0.5	0.984 (0.004)	0.045 (0.001)	0.019 (0.004)	0.984 (0.007)	0.045 (0.001)	0.022 (0.006)
0.75	0.990 (0.004)	0.047 (0.001)	0.014 (0.006)	0.985 (0.007)	0.050 (0.001)	0.016 (0.006)
0.9	0.987 (0.009)	0.049 (0.001)	0.012 (0.009)	0.984 (0.008)	0.052 (0.001)	0.002 (0.012)
0.97	0.980 (0.014)	0.050 (0.002)	0.023 (0.015)	0.982 (0.013)	0.053 (0.001)	0.021 (0.018)

Transformation to Normality

A presumed advantage of univariate (age-specific) transformation to normality is that once observations are transformed to univariate "Z-scores" they are automatically prepared to longitudinal autoregression:

$$Z_t = \alpha_0 + \alpha_1 Z_{t-1} + U_t$$

Premise: Marginal Normality $\Rightarrow$ Joint Normality

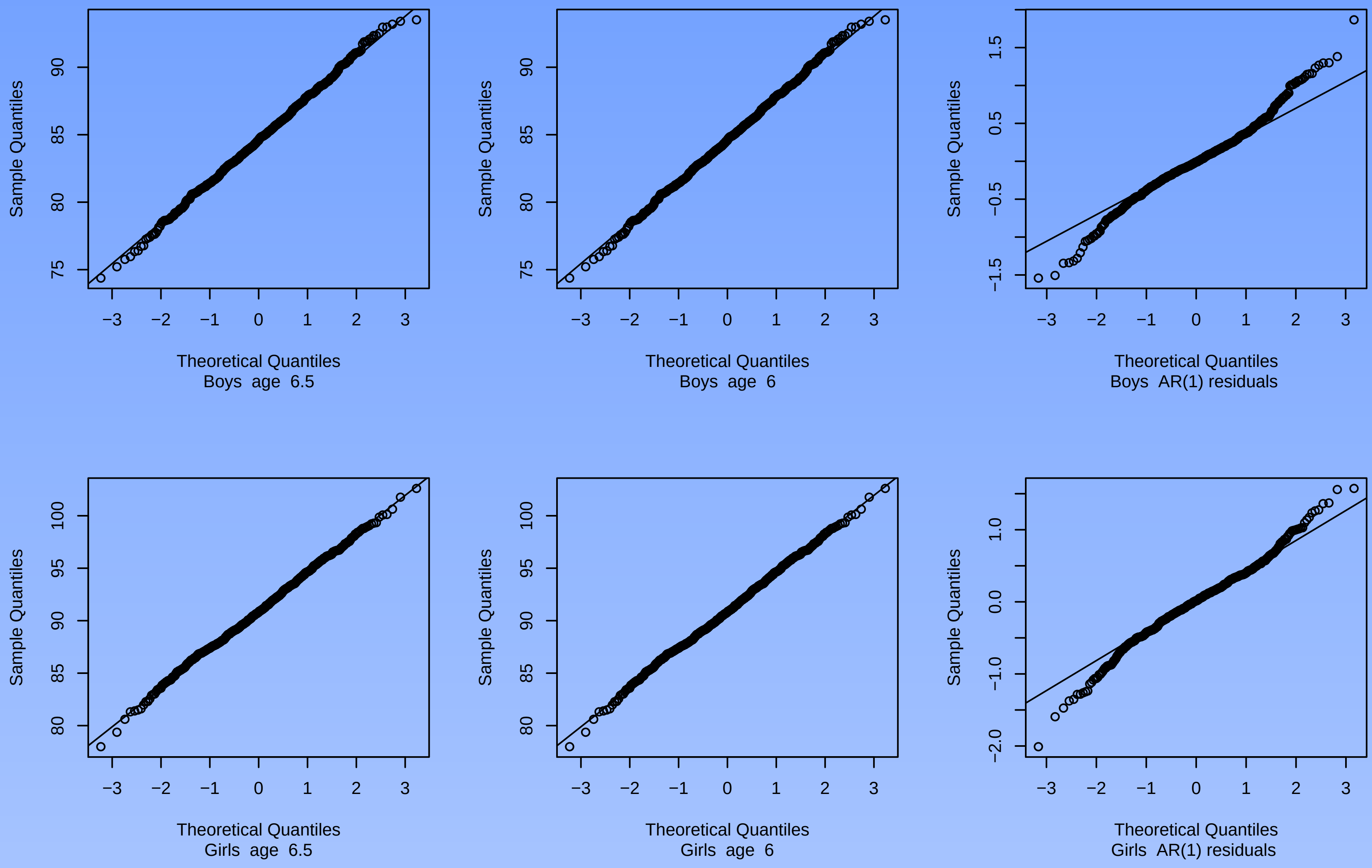
Transformation to Normality

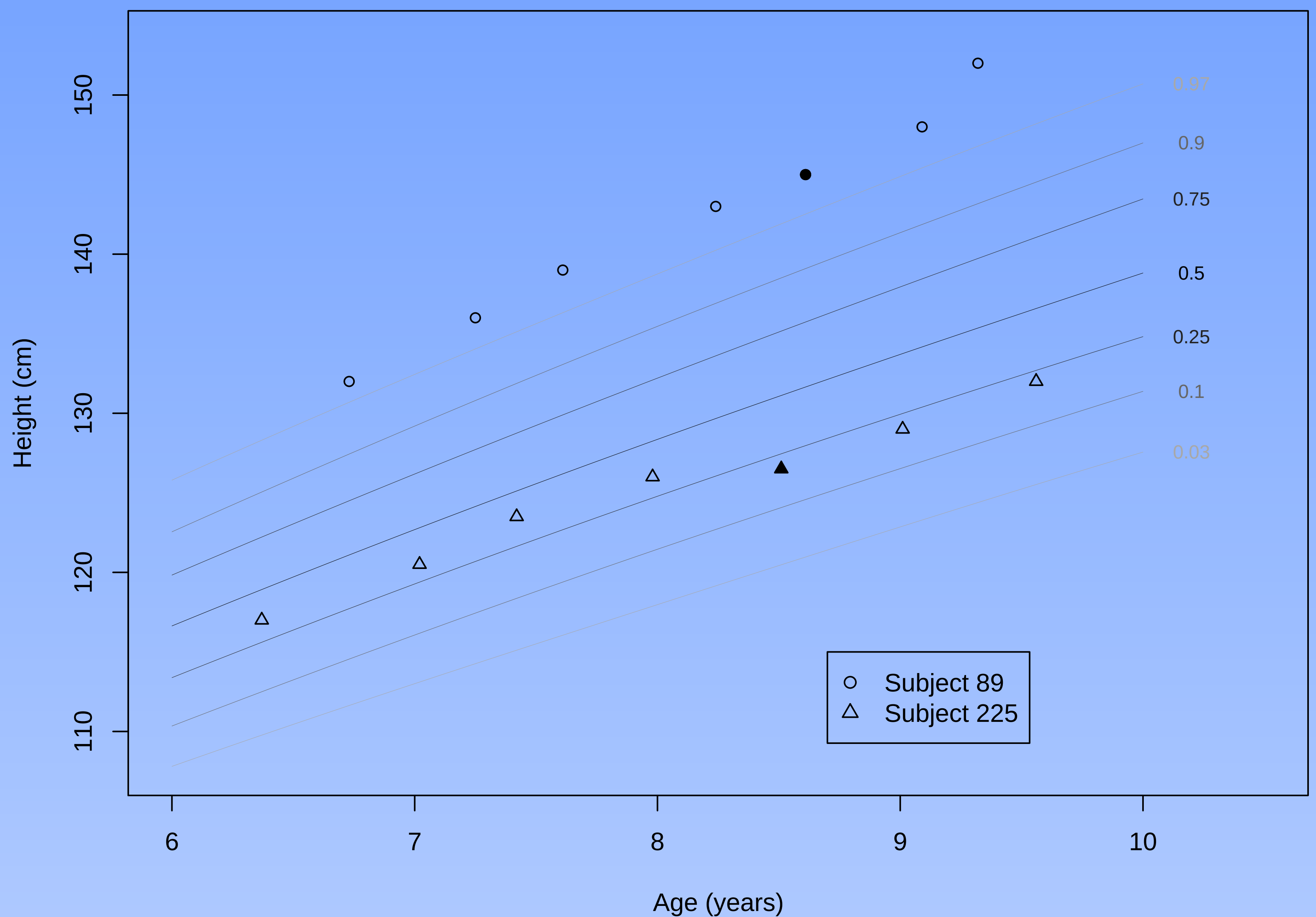
A presumed advantage of univariate (age-specific) transformation to normality is that once observations are transformed to univariate "Z-scores" they are automatically prepared to longitudinal autoregression:

$$Z_t = \alpha_0 + \alpha_1 Z_{t-1} + U_t$$

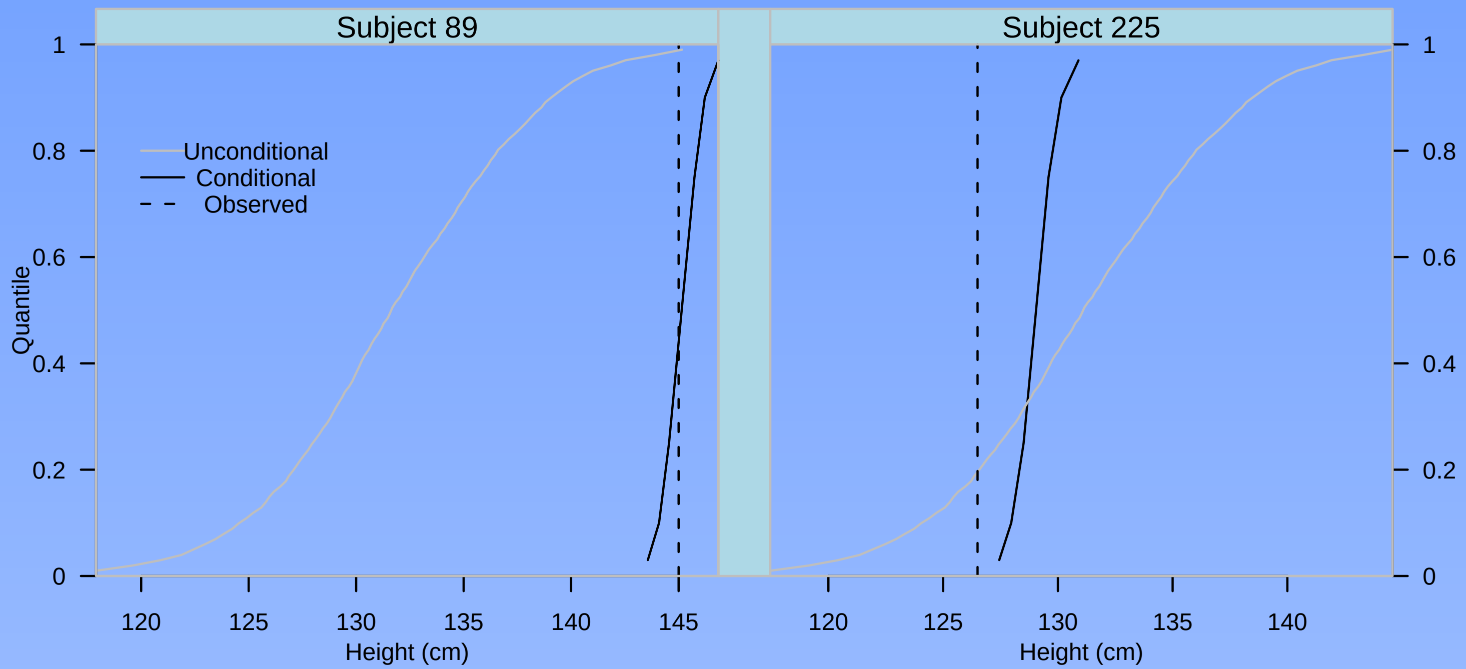
Premise: Marginal Normality $\Rightarrow$ Joint Normality

Of course we know it isn't true, but we also think we know that counterexamples are pathological, and don't occur in "nature."


Sample Quantiles
Theoretical Quantiles
Boys age 6.5
Sample Quantiles
Theoretical Quantiles
Boys age 6
Sample Quantiles
Theoretical Quantiles
Boys AR(1) residuals
Sample Quantiles
Theoretical Quantiles
Girls age 6.5
Sample Quantiles
Theoretical Quantiles
Girls age 6
Sample Quantiles
Theoretical Quantiles
Girls AR(1) residuals



Height (cm)
110
120
130
140
150
Age (years)
6
7
8
9
10
0.97
0.9
0.75
0.5
0.25
0.1
0.03
Subject 89
Subject 225


Subject 89
Subject 225
Unconditional
Conditional
Observed
Quantile
1
0.8
0.6
0.4
0.2
0
120
125
130
135
140
145
Height (cm)
120
125
130
135
140
Height (cm)

Conclusions

- Nonparametric quantile regression using B-splines offers a reasonable alternative to parametric methods for constructing reference growth charts.

Conclusions

- Nonparametric quantile regression using B-splines offers a reasonable alternative to parametric methods for constructing reference growth charts.

- The flexibility of quantile regression methods exposes features of the data that are not easily observable with conventional parametric methods. *Even for height data.*

Conclusions

- Nonparametric quantile regression using B-splines offers a reasonable alternative to parametric methods for constructing reference growth charts.
- The flexibility of quantile regression methods exposes features of the data that are not easily observable with conventional parametric methods. *Even for height data.*
- Longitudinal data can be easily accomodated into the quantile regression framework by adding covariates, including the use of autoregressive effects for unequally spaced measurements.