Beyond the Average Man:
The Empirics of Heterogeneity in Social Science

Roger Koenker
University of Illinois at Urbana-Champaign

University of Northern Arizona: October 21, 2005

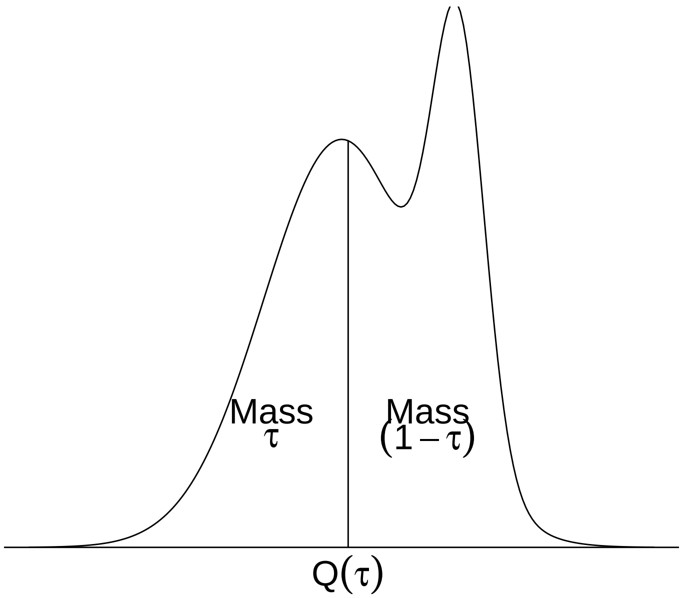
Talk based on joint work with Kevin Hallock (Cornell U.), Ying Wei (Columbia U.) Anneli Pere (U. of Helsinki), Xuming He (UIUC), and Pin Ng (NAU). All computations and graphics were done in the R language; slides were produced by `pdflatex | ppower4`.

Adolphe Quetelet (1796 - 1874)

Father of Quantitative Social Social Science.

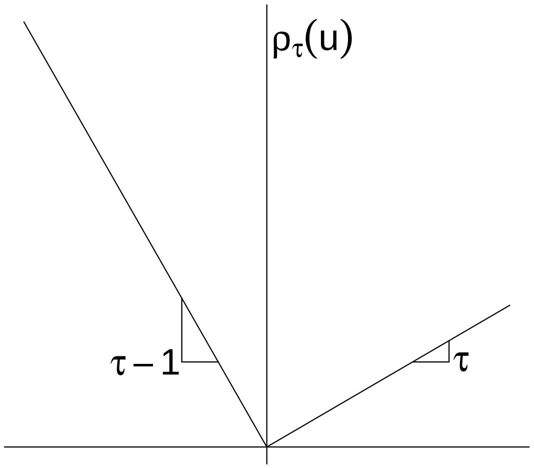
Discoverer of the Average Man.

Quantiles



The τth quantile, $Q(\tau)$, divides a univariate distribution into two parts: mass to the left of $Q(\tau)$ is τ, mass to the right of $Q(\tau)$ is $(1 - \tau)$.

Sample Quantiles via Optimization



$$Q(\tau) = \text{argmin}_{a \in \Re} \int \rho_\tau(y - a)dF(y)$$

$$\hat{Q}(\tau) = \text{argmin}_{a \in \Re} \sum_{i=1}^{n} \rho_\tau(y_i - a)$$

How/Why Does It Work?

Median: Minimizers must balance the mass (number of observations) above and below the estimate so that they are equal.

Quantiles: Minimizers must asymmetrically balance the mass so that,

$$\tau\{\#\text{Above}\} + (\tau - 1)\{\#\text{Below}\} = 0$$

This requires that $\{\#\text{Above}\}$ be roughly $(1 - \tau)n$ and $\{\#\text{Below}\}$ be roughly τn

How can these ideas be extended to the regression setting?

The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m E(Y - m)^2$$

The conditional mean $\mu(x) = E(Y|X = x)$ solves

$$\mu(x) = \min_m E_{Y|X=x}(Y - m(X))^2.$$

Similarly, the unconditional τth quantile solves

$$\alpha_\tau = \min_a E\rho_\tau(Y - a)$$

and the conditional τth quantile solves

$$\alpha_\tau(x) = \min_a E_{Y|X=x}\rho_\tau(Y - a(X))$$

Regression Quantiles via Optimization

The sample analogue of the foregoing population concepts yields, the nonparametric quantile regression estimator

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

If we take $\mathcal{A} = \{a : \mathbb{R}^p \to \mathbb{R} | a(x) = x^\top \beta,\ \beta \in \mathbb{R}^p\}$, then we have the linear (in parameters) quantile regression problem:

$$\hat{\beta}(\tau) = \text{argmin}_{b \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b)$$

Computation of Quantile Regression

Primal Formulation as a Linear Program

$$\min\{\tau 1^\top u + (1-\tau)1^\top v | y = Xb + u - v, (b,u,v) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}\}$$

Dual Formulation as a Linear Program

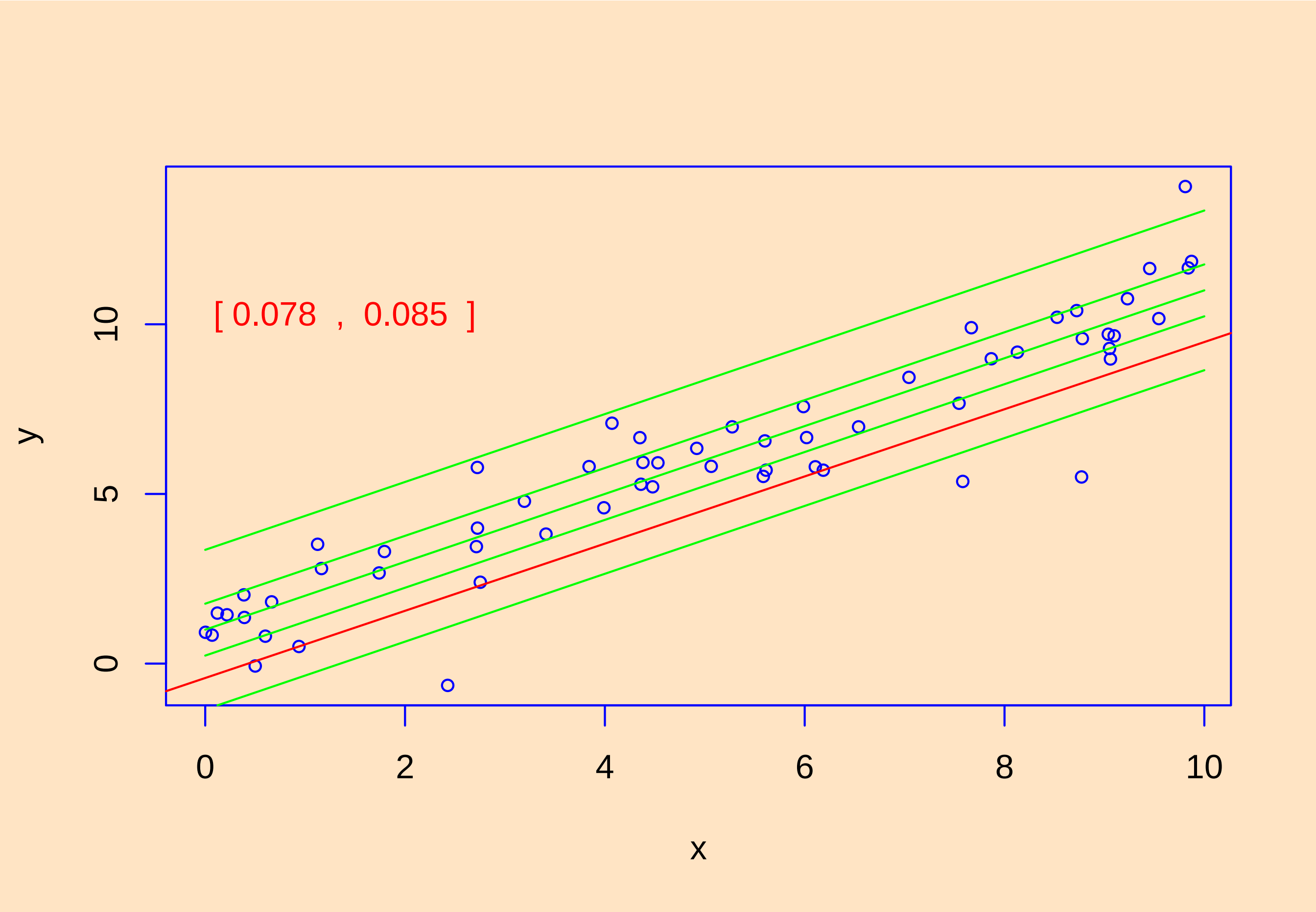
$$\max\{y'd | X^\top d = (1-\tau)X^\top 1, d \in [0,1]^n\}$$

Solutions are characterized by an exact fit to p observations.

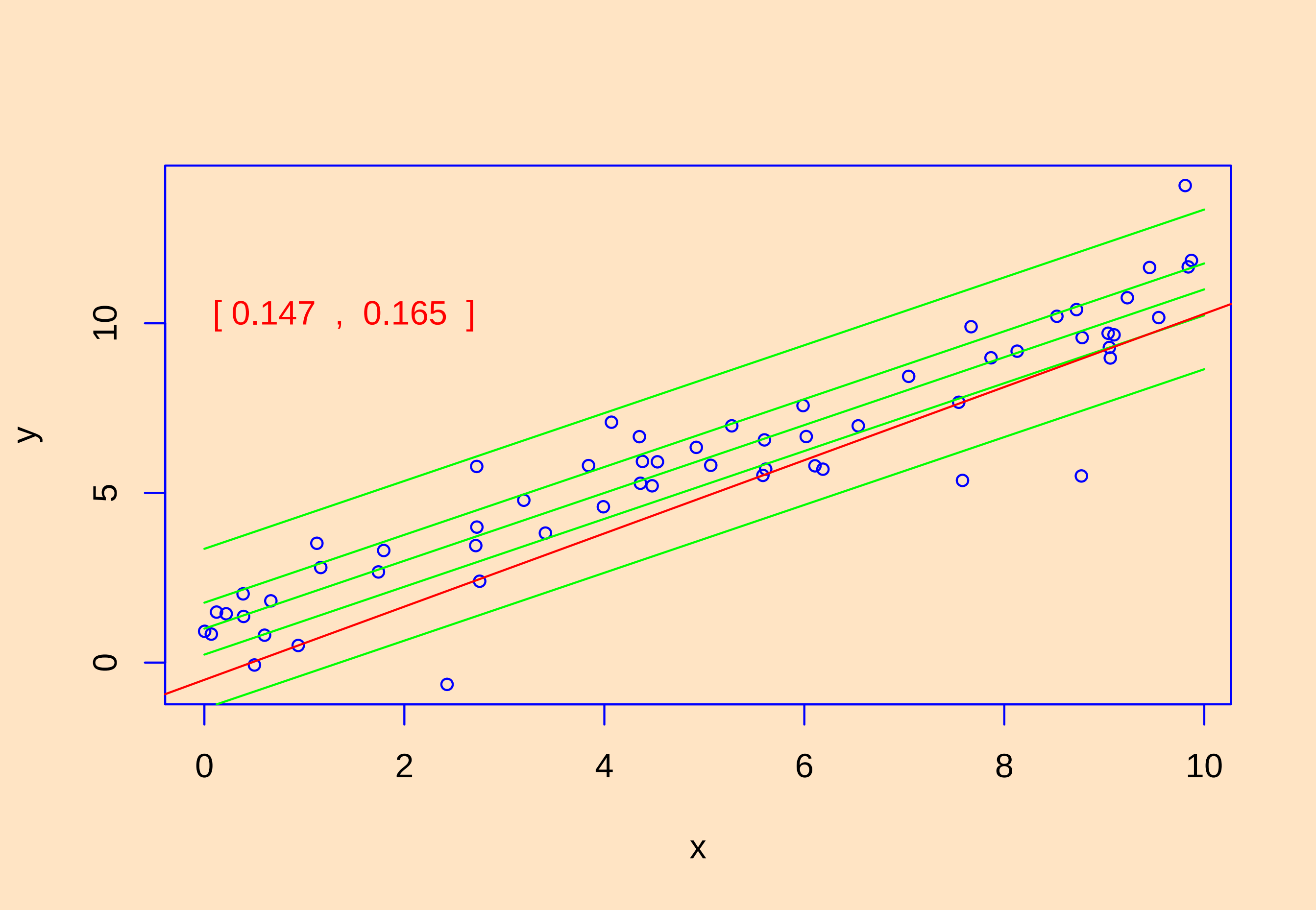
Quantile Regression: A Movie

- Bivariate linear model with iid Student t errors
- Conditional quantile functions are parallel in green
- 100 observations indicated in blue
- Fitted quantile regression lines in red

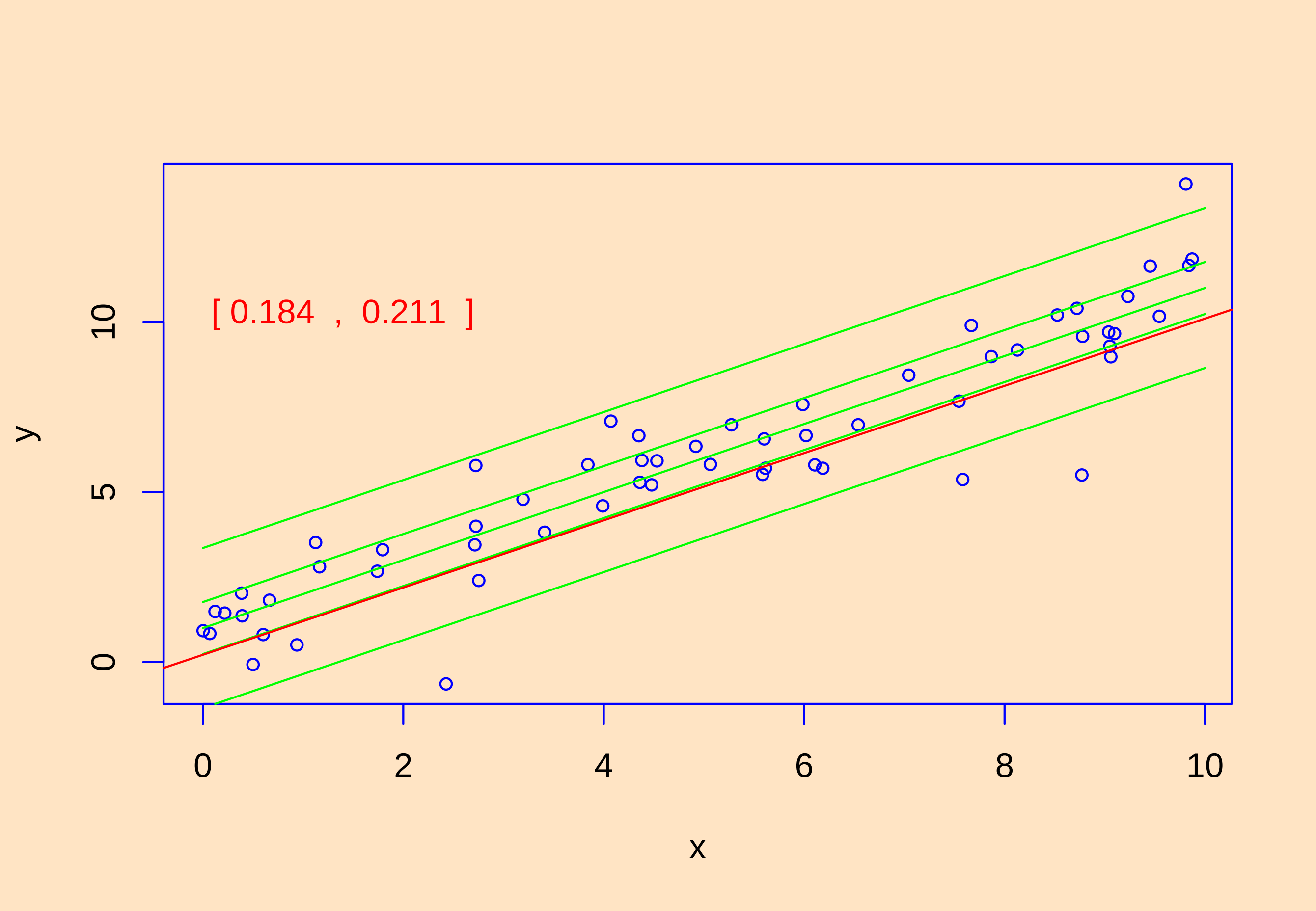
Quantile Regression in the iid Error Model



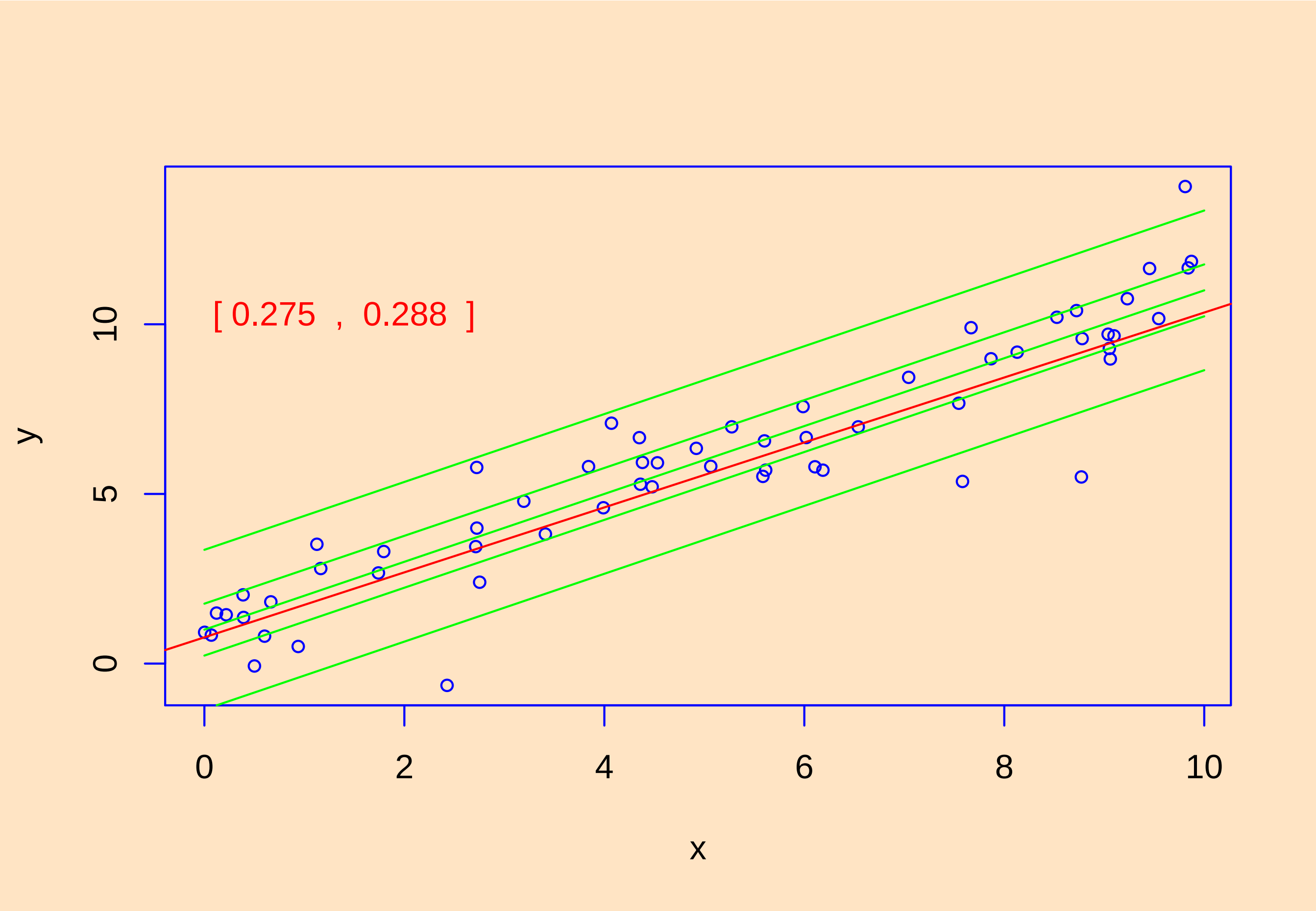
Quantile Regression in the iid Error Model



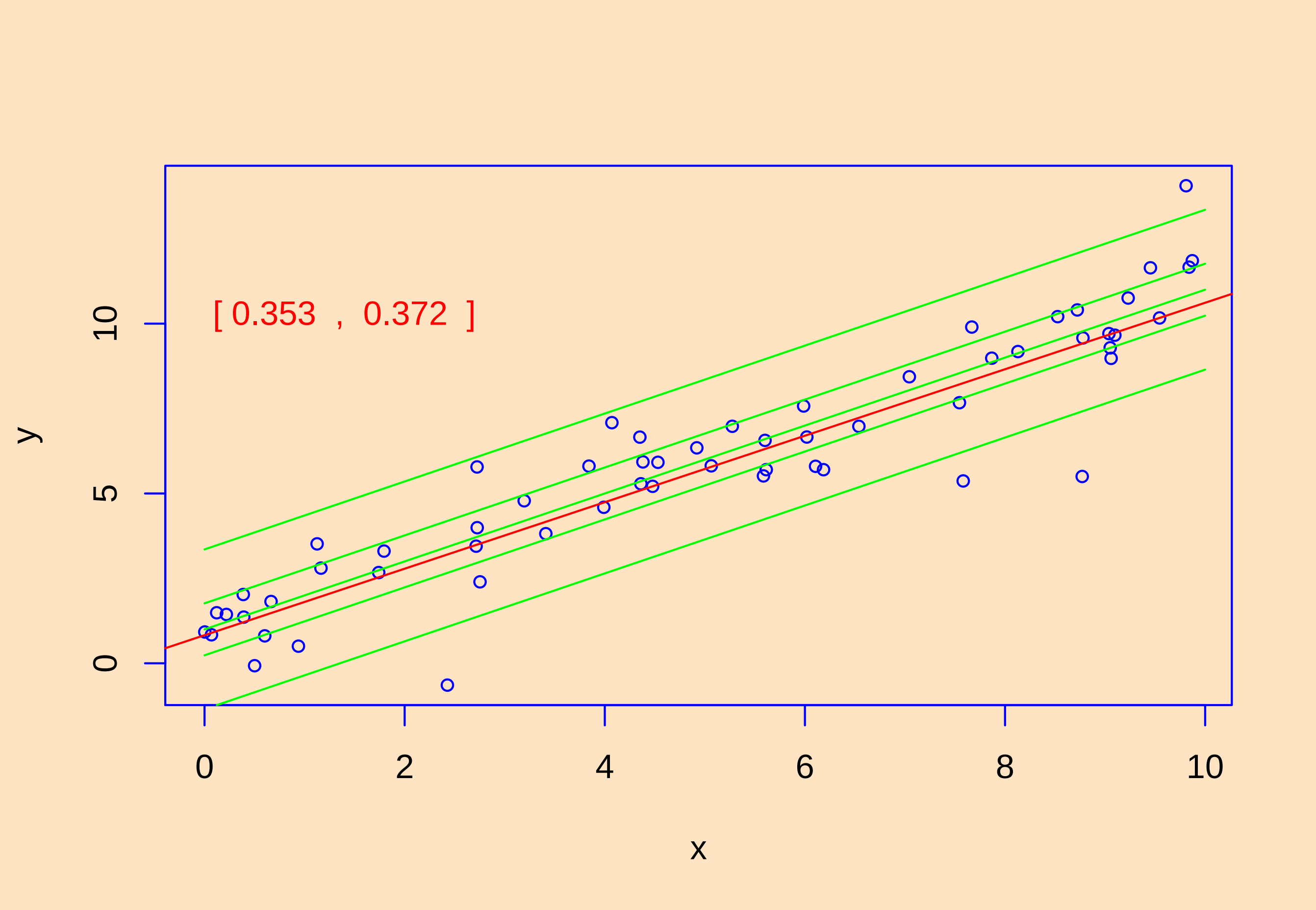
Quantile Regression in the iid Error Model



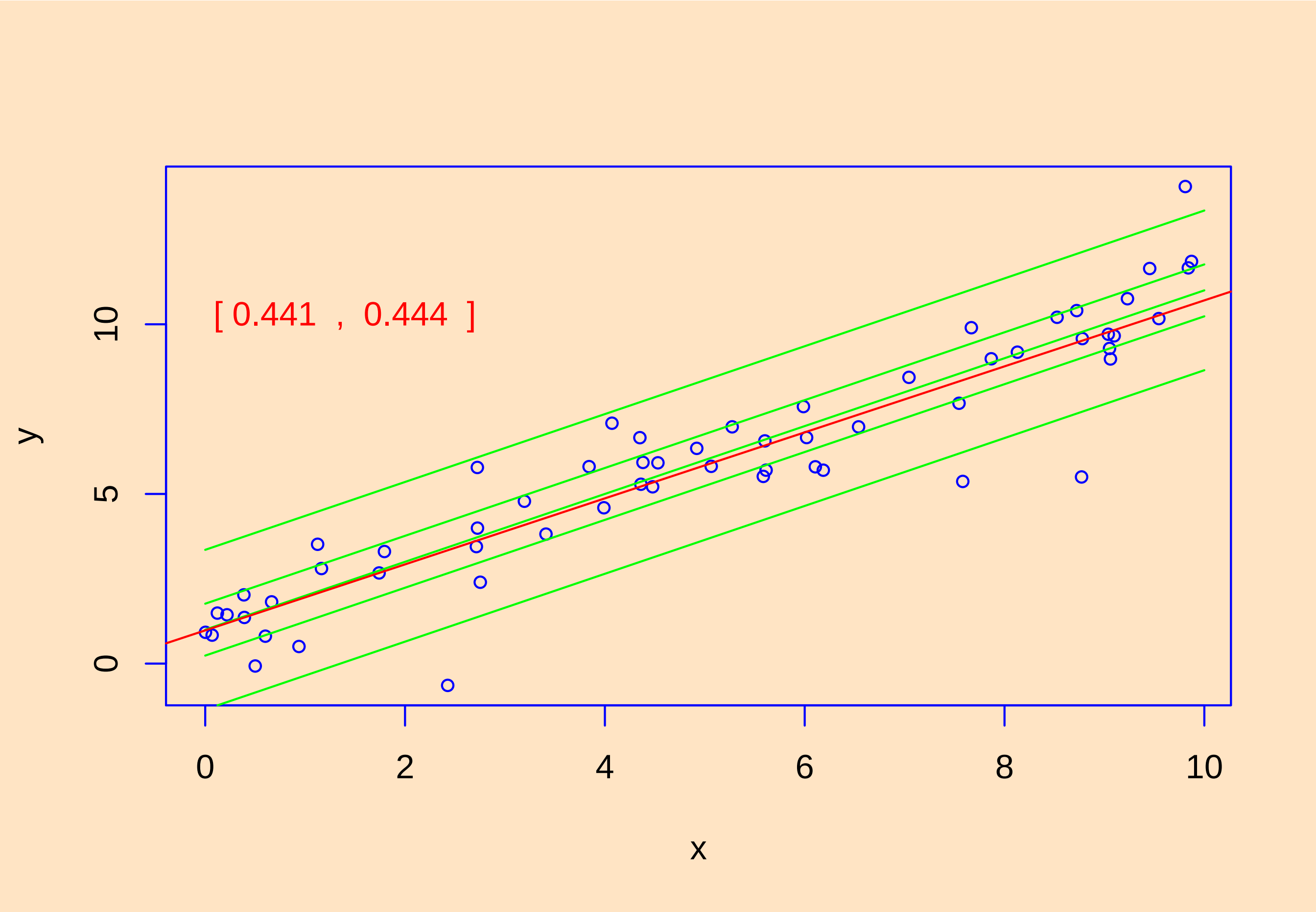
Quantile Regression in the iid Error Model



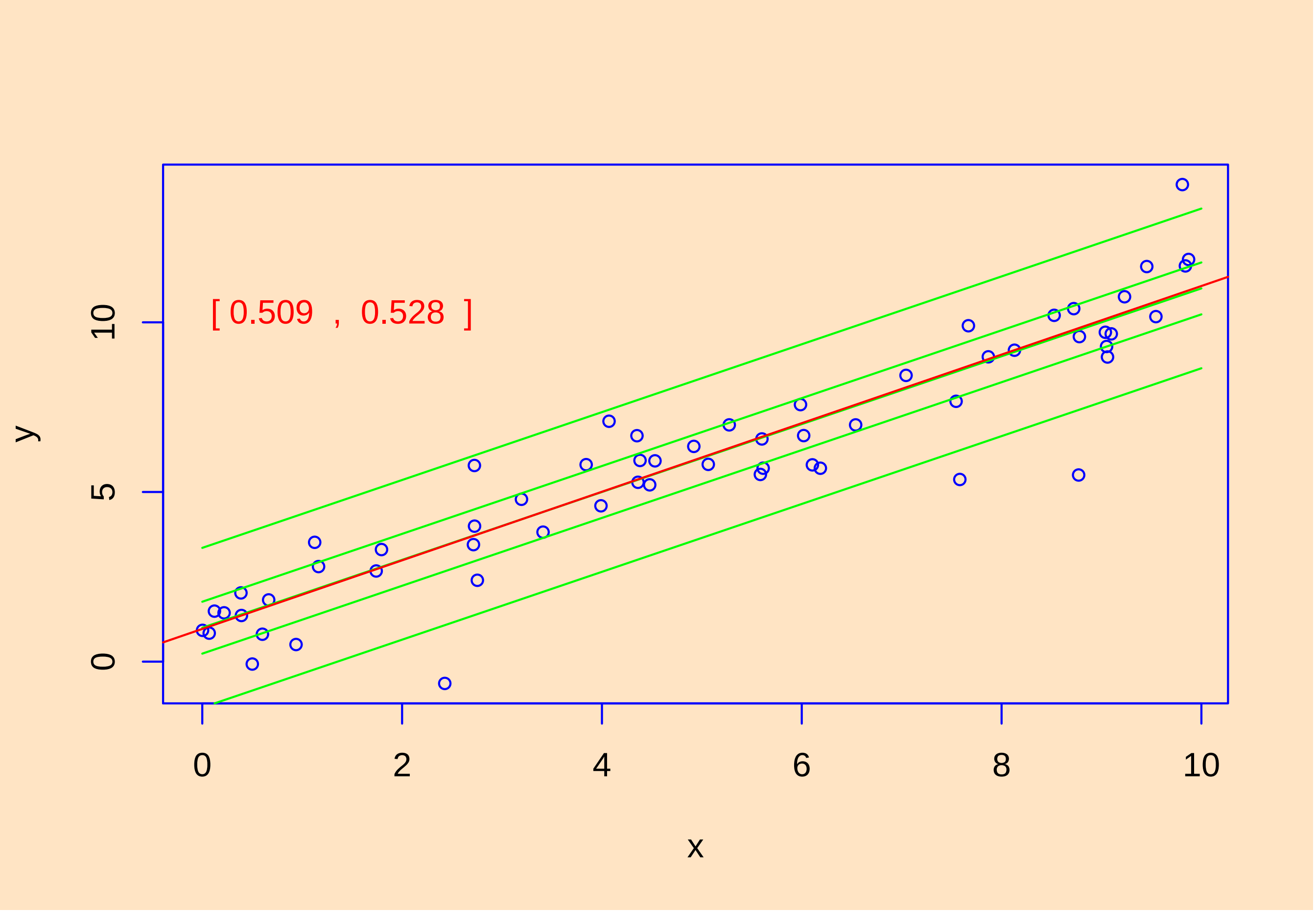
Quantile Regression in the iid Error Model



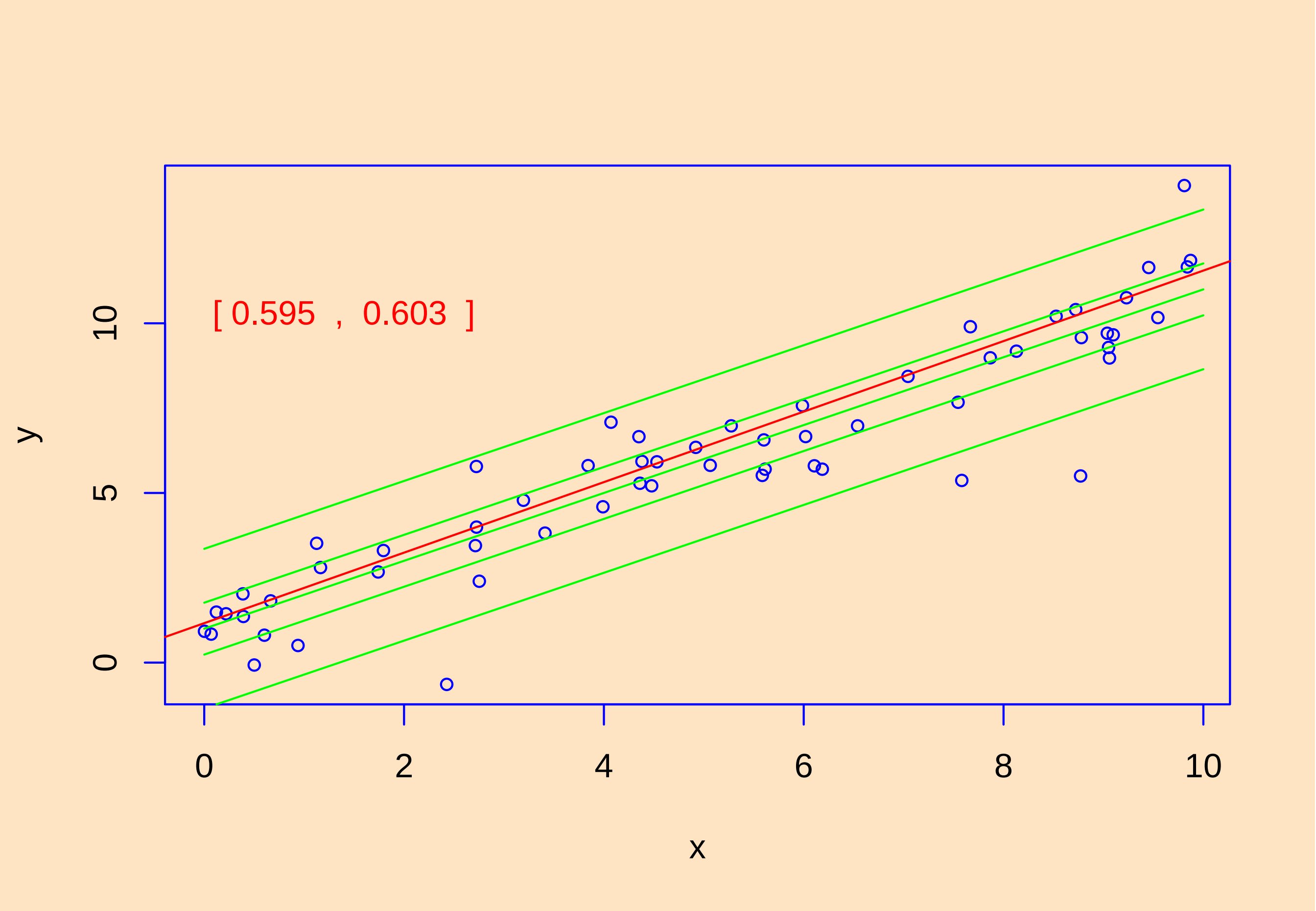
Quantile Regression in the iid Error Model



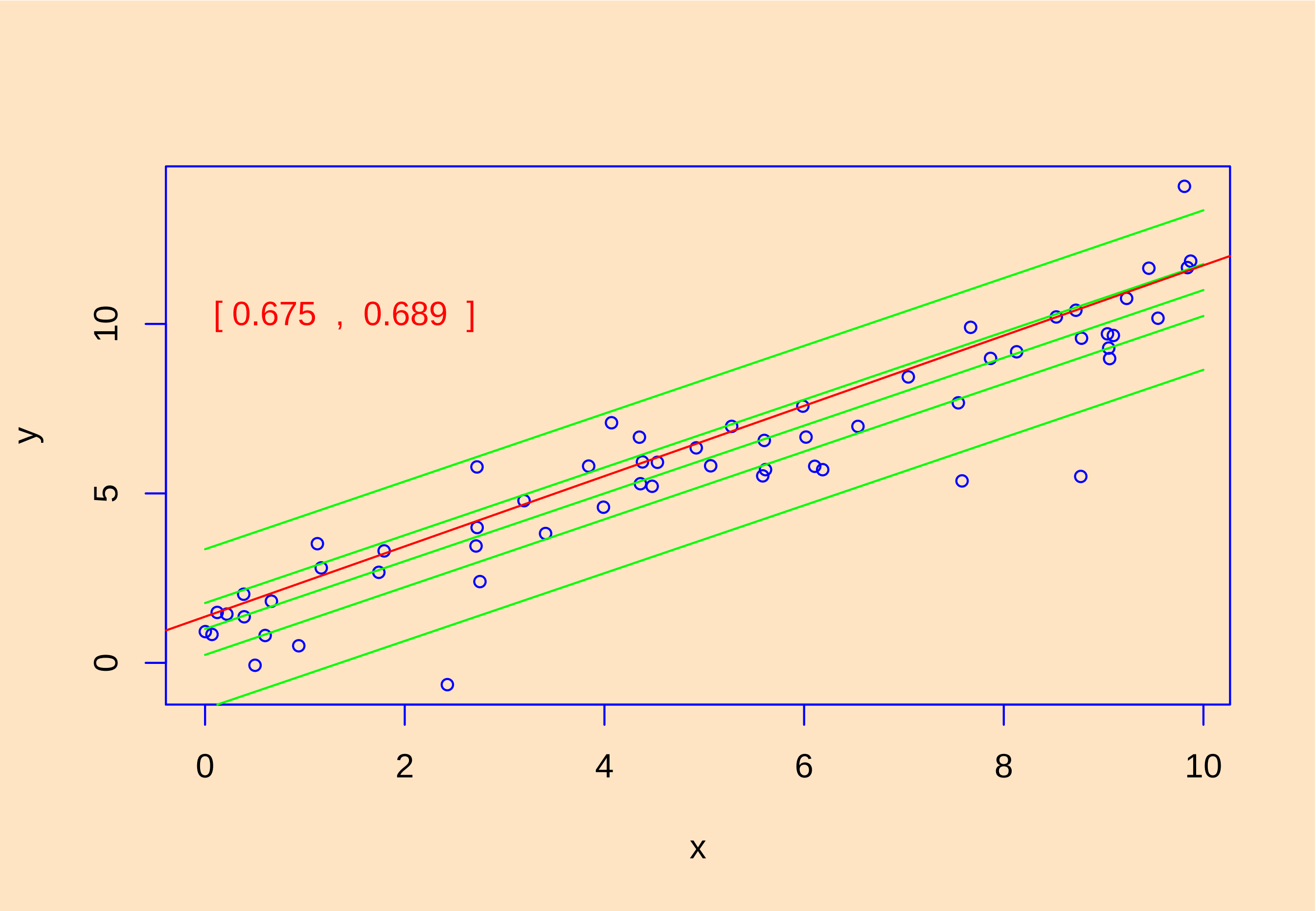
Quantile Regression in the iid Error Model



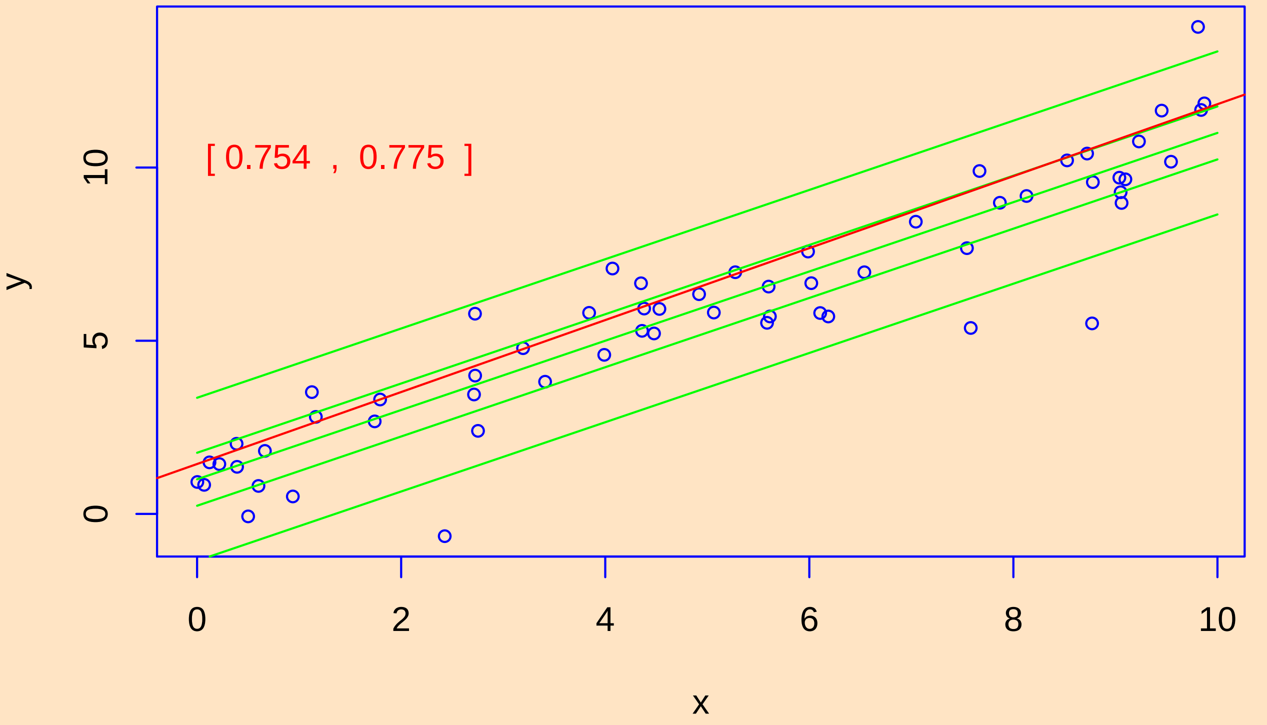
Quantile Regression in the iid Error Model



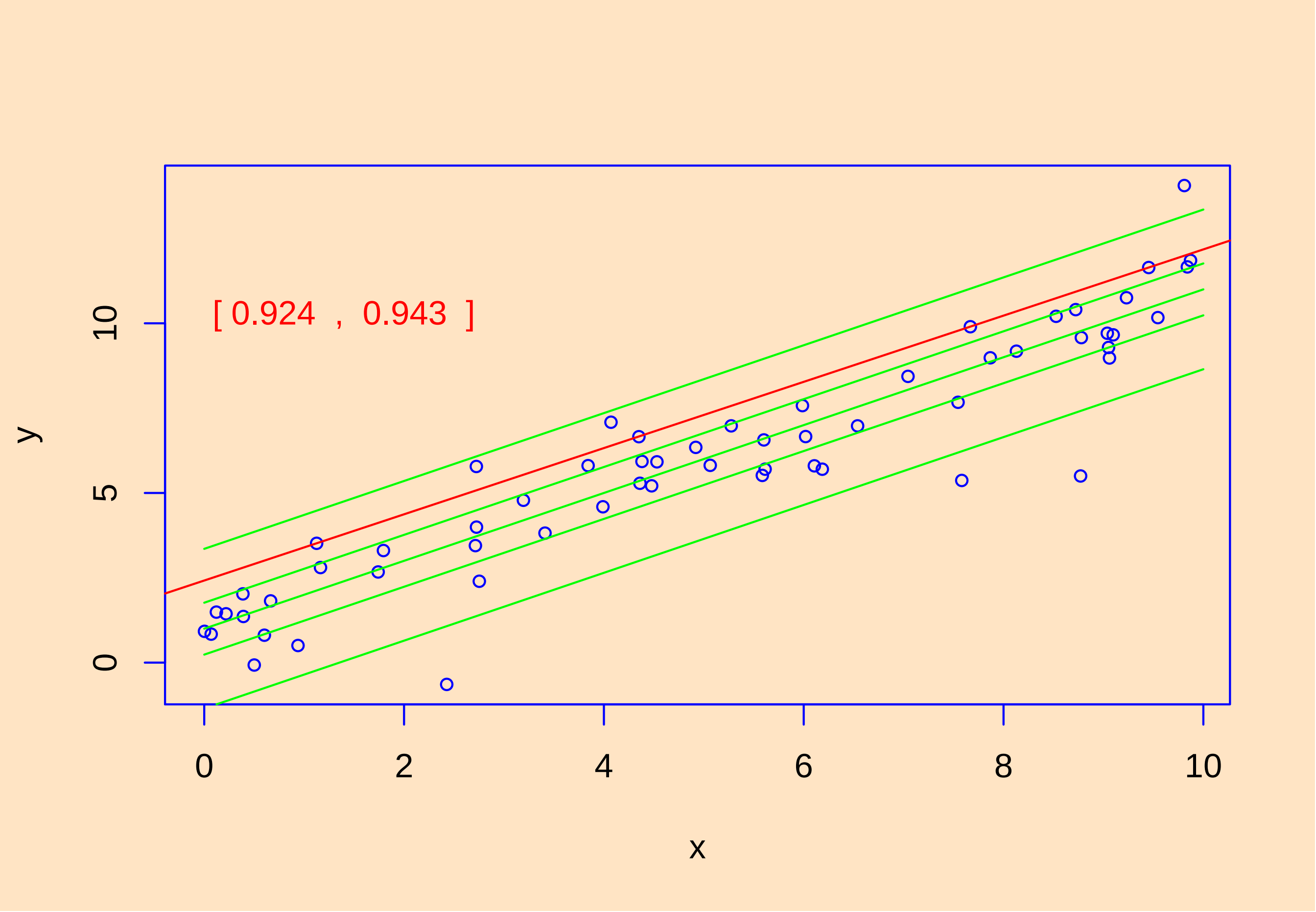
Quantile Regression in the iid Error Model



Quantile Regression in the iid Error Model



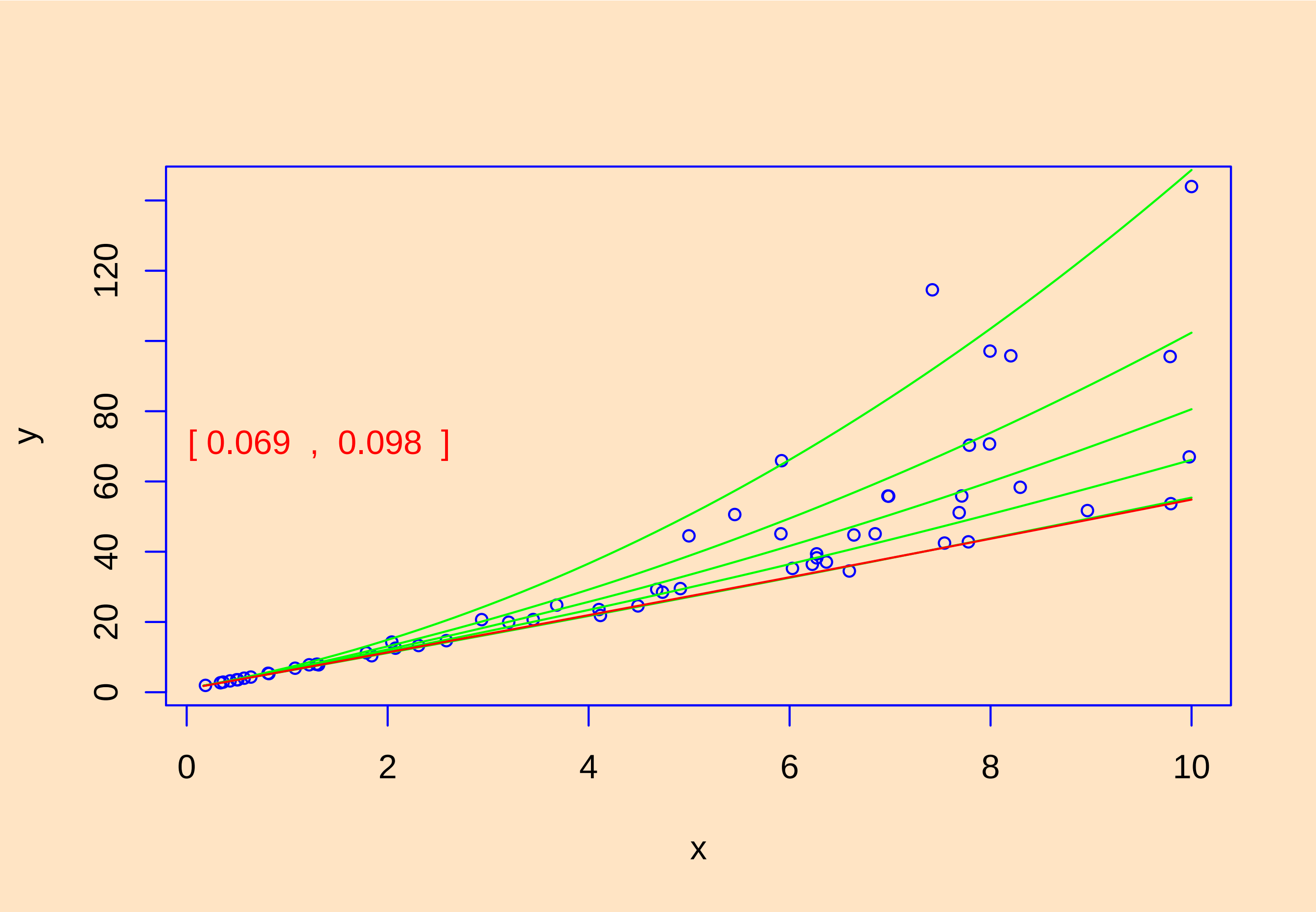
Quantile Regression in the iid Error Model



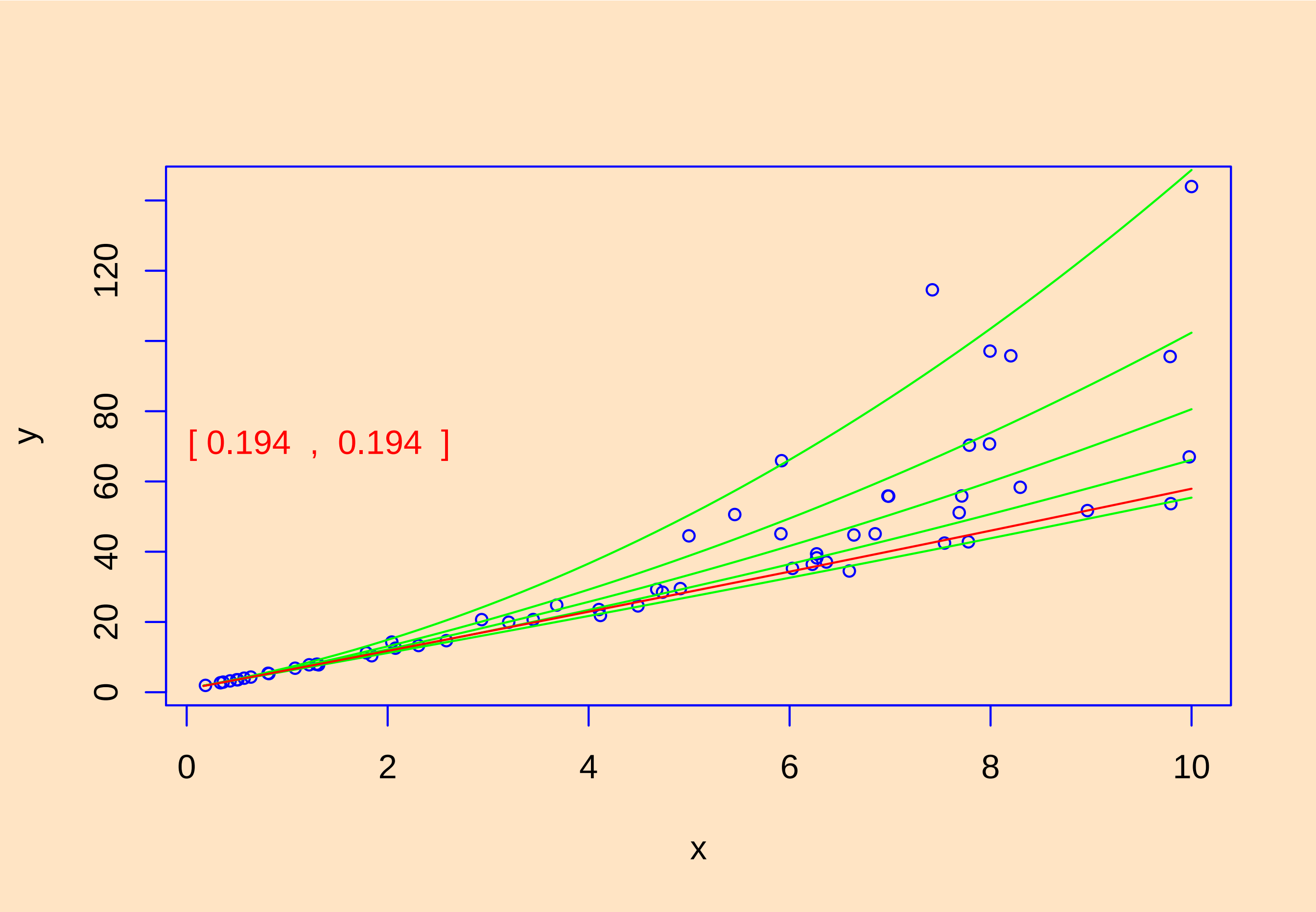
Virtual Quantile Regression II

- Bivariate quadratic model with Heteroscedastic χ^2 errors
- Conditional quantile functions drawn in green
- 100 observations indicated in blue
- Fitted quadratic quantile regression lines in red

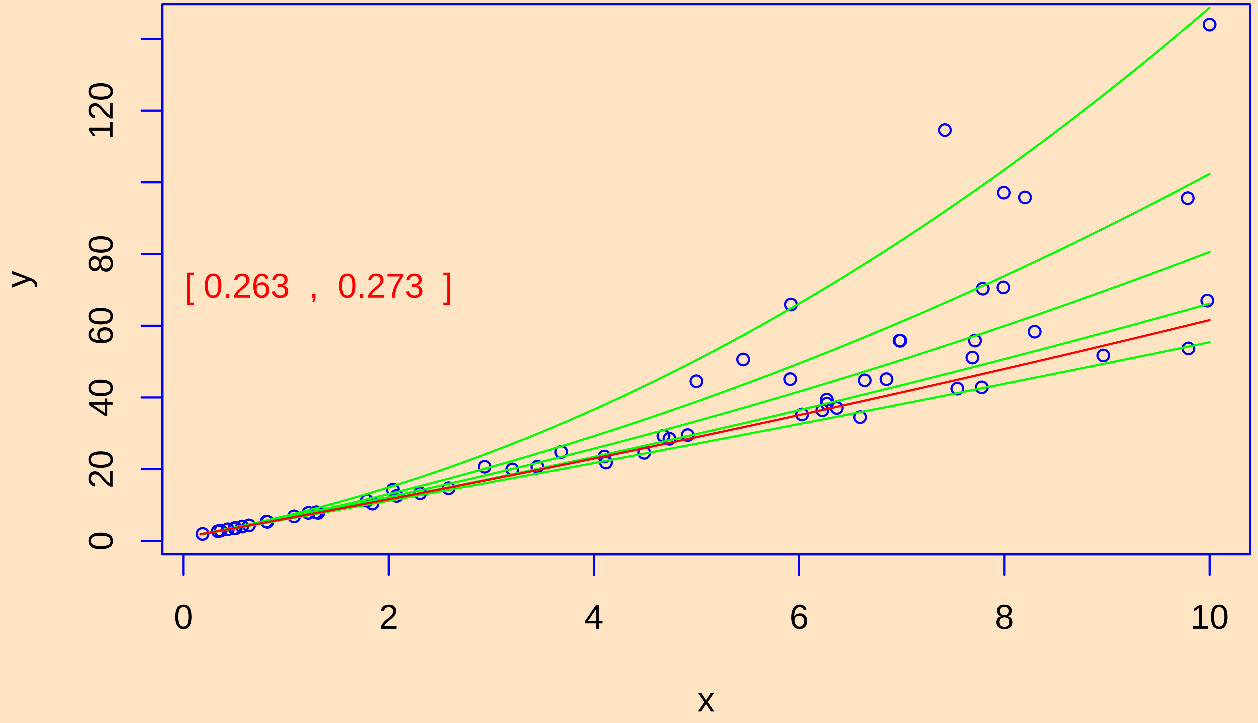
Quantile Regression in the Heteroscedastic Error Model



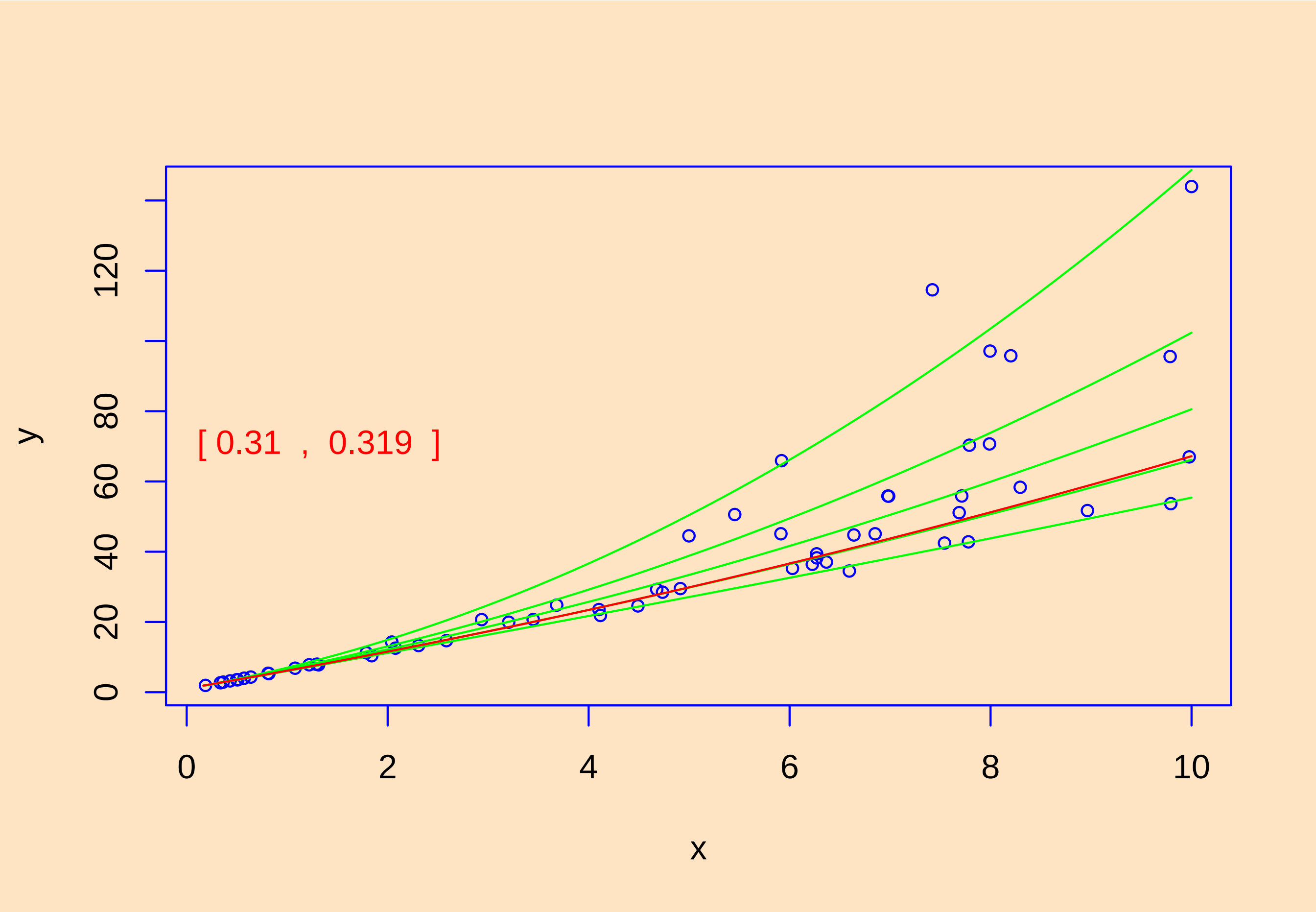
Quantile Regression in the Heteroscedastic Error Model



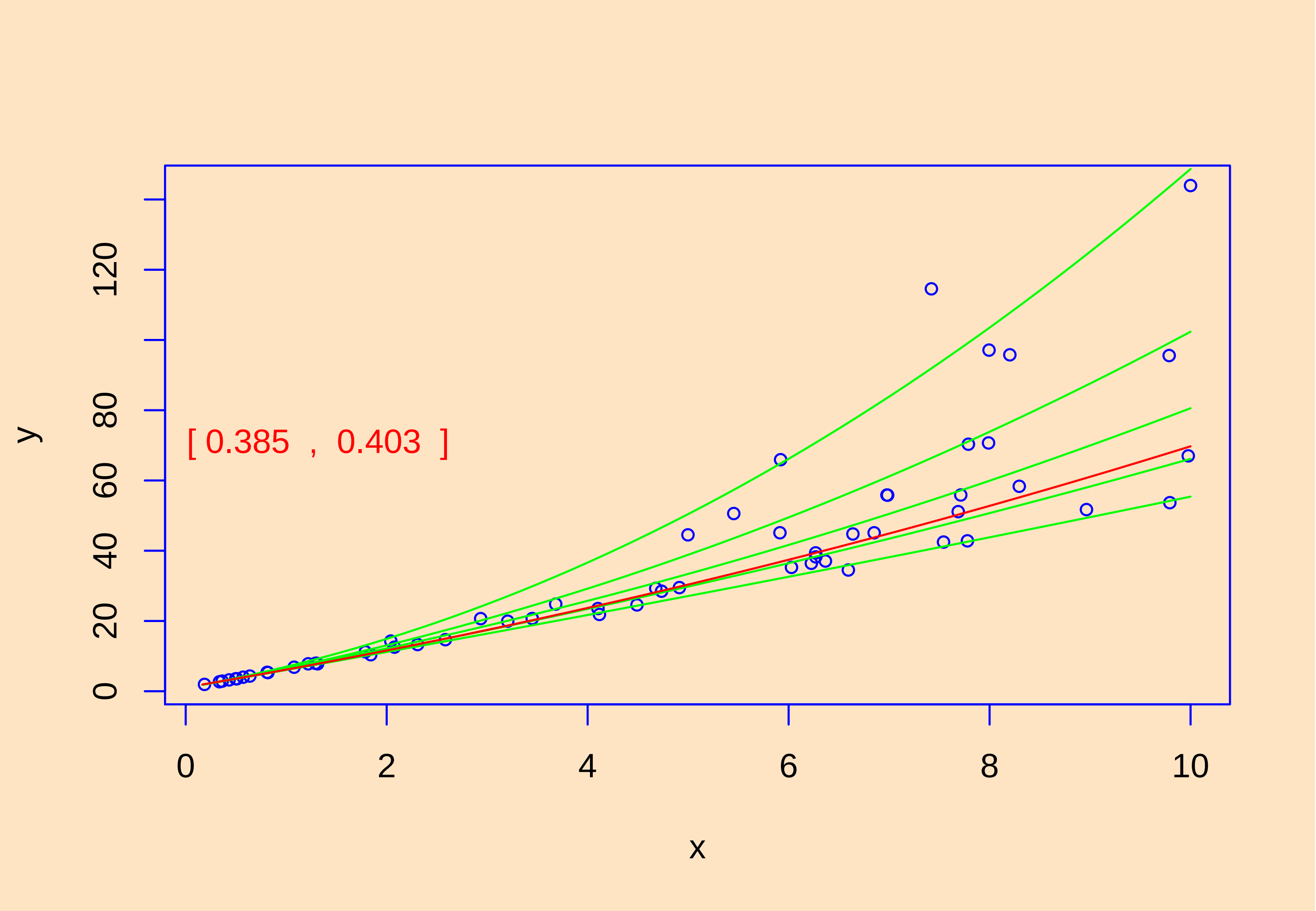
Quantile Regression in the Heteroscedastic Error Model



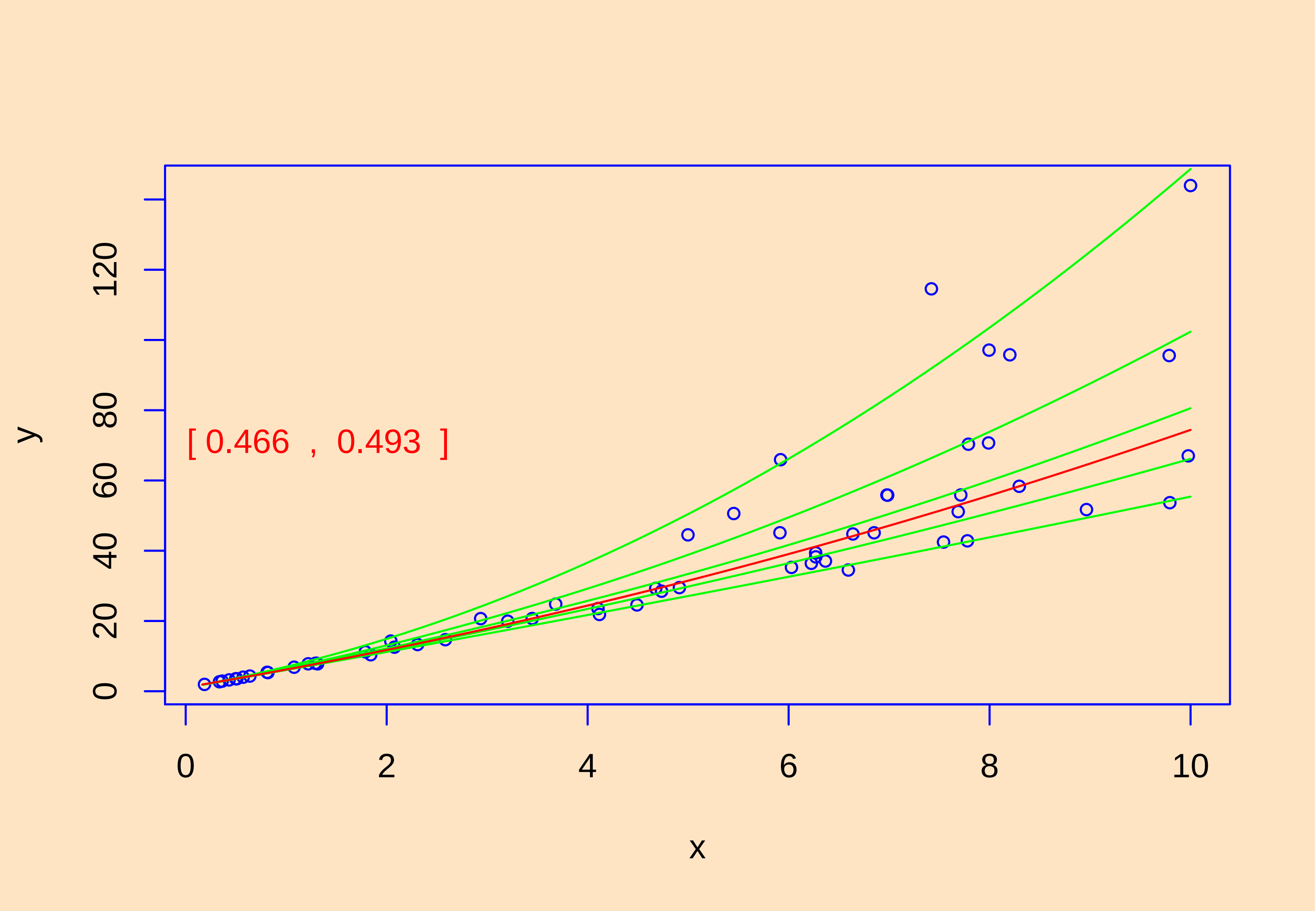
Quantile Regression in the Heteroscedastic Error Model



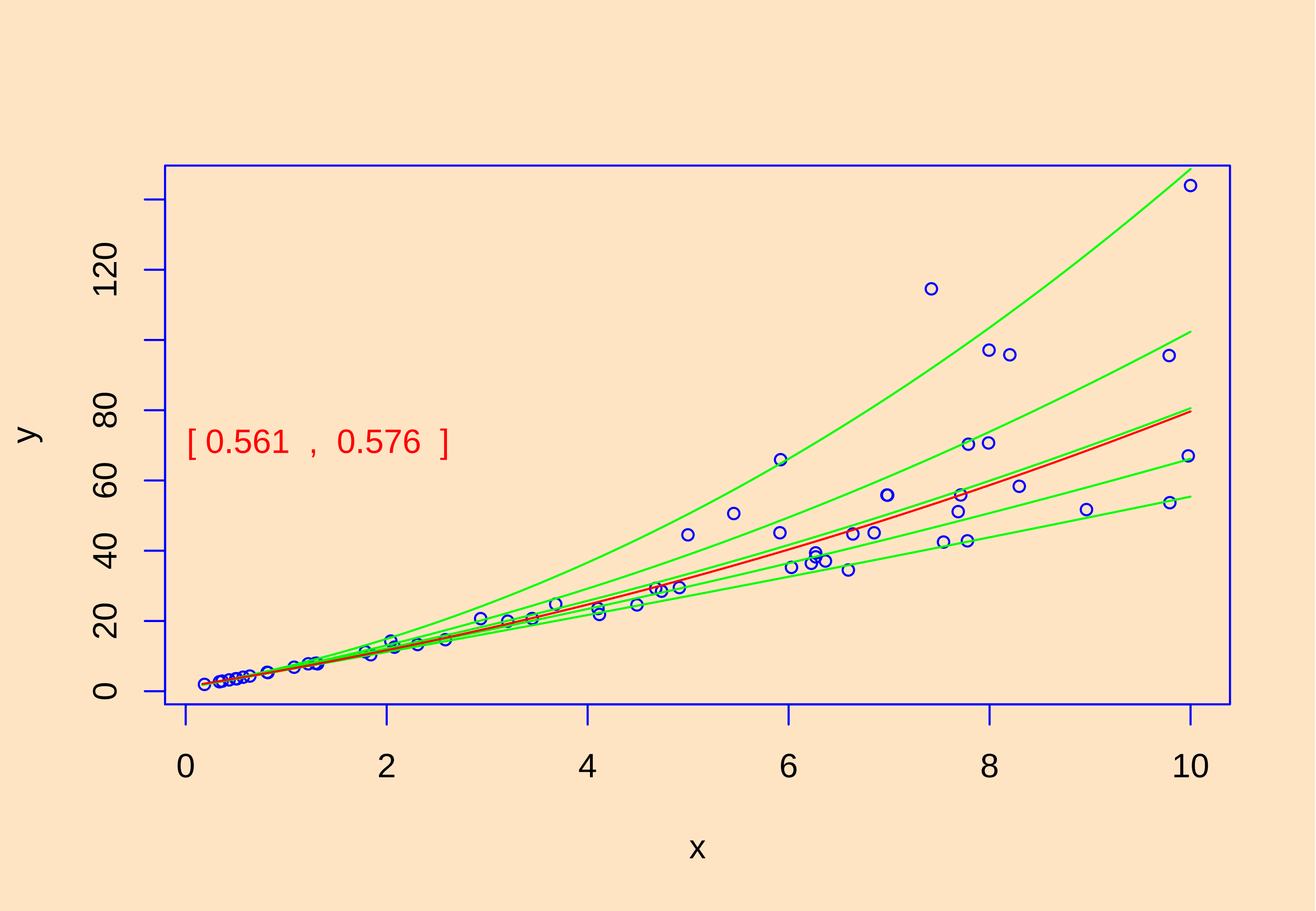
Quantile Regression in the Heteroscedastic Error Model



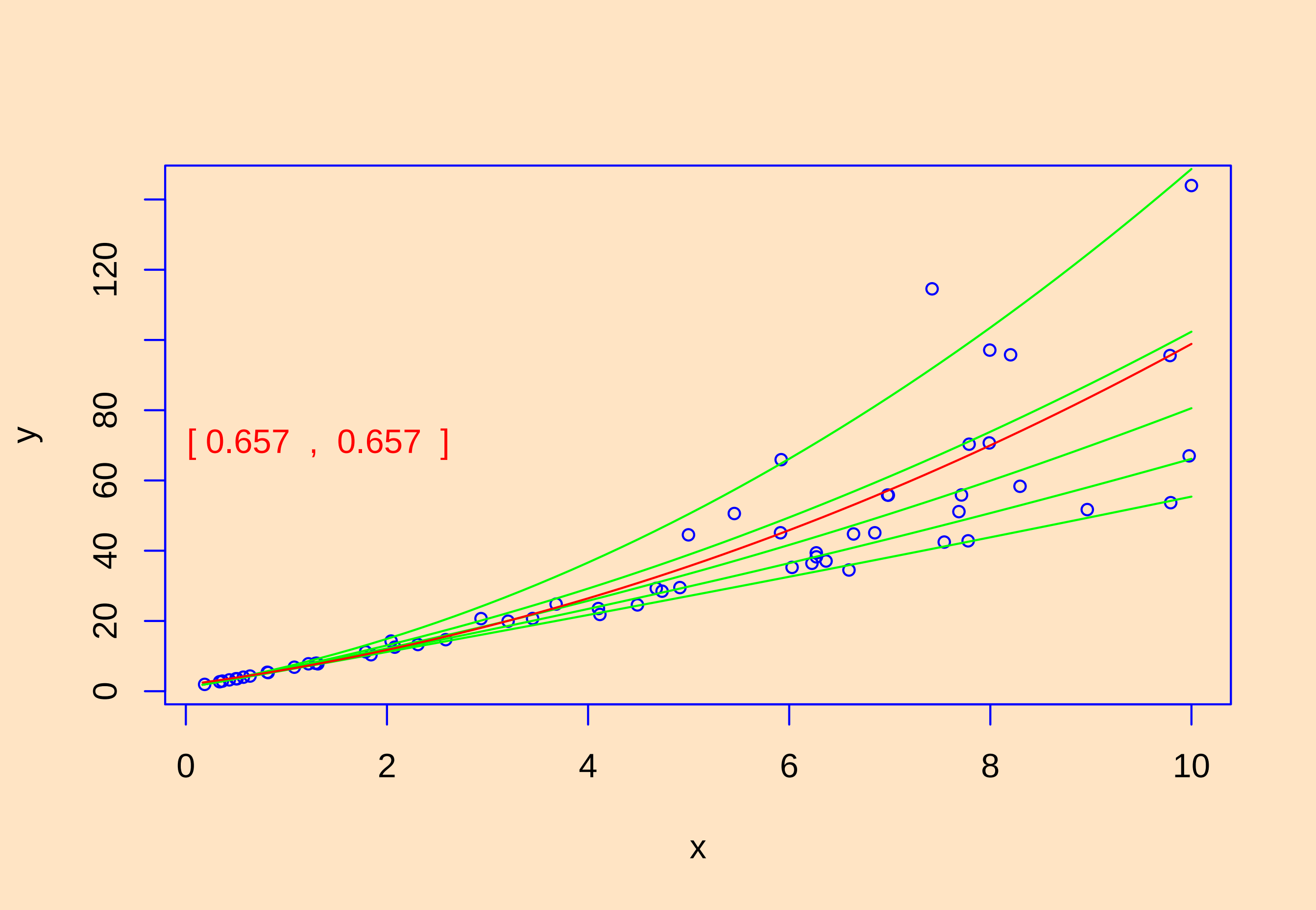
Quantile Regression in the Heteroscedastic Error Model



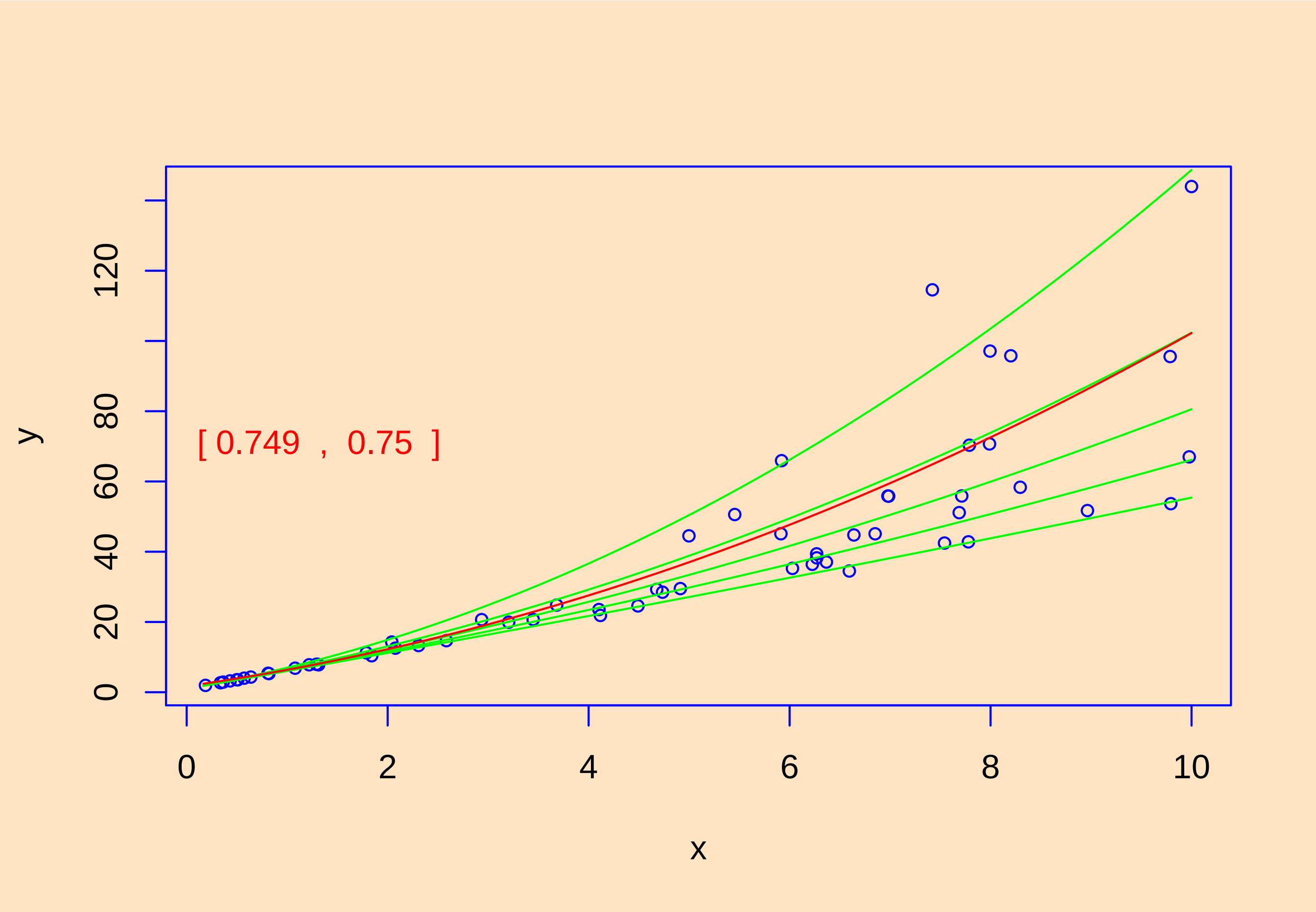
Quantile Regression in the Heteroscedastic Error Model



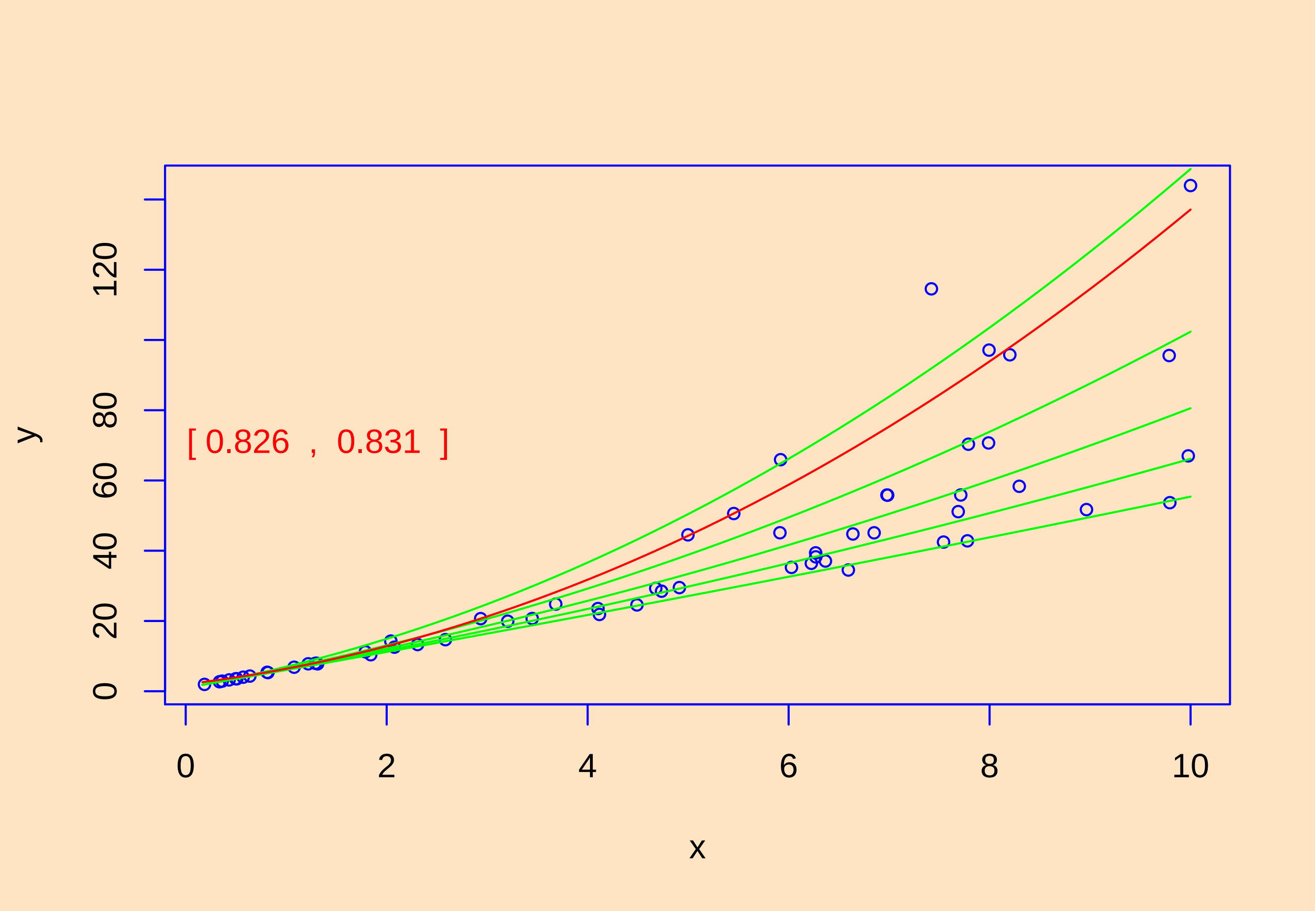
Quantile Regression in the Heteroscedastic Error Model



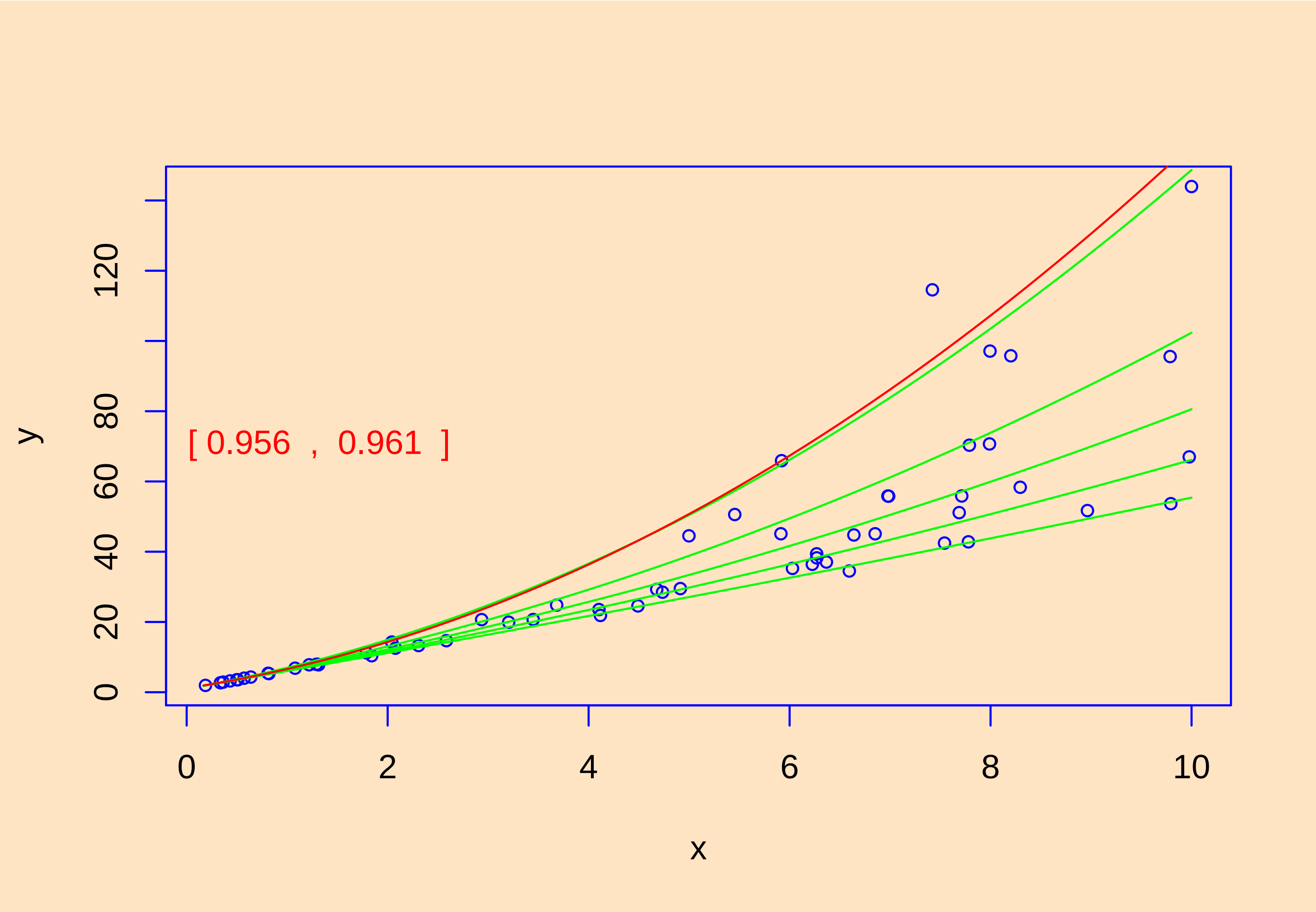
Quantile Regression in the Heteroscedastic Error Model



Quantile Regression in the Heteroscedastic Error Model



Quantile Regression in the Heteroscedastic Error Model



Beyond Average Treatment Effects

Lehmann (1974) proposed the following general model of treatment response:

> "Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment responses is that of the random variable $X + \Delta(X)$ where X is distributed according to F."

Lehmann QTE as a QQ-Plot

Doksum (1974) defines $\Delta(x)$ as the "horizontal distance" between F and G at x, *i.e.*

$$F(x) = G(x + \Delta(x)).$$

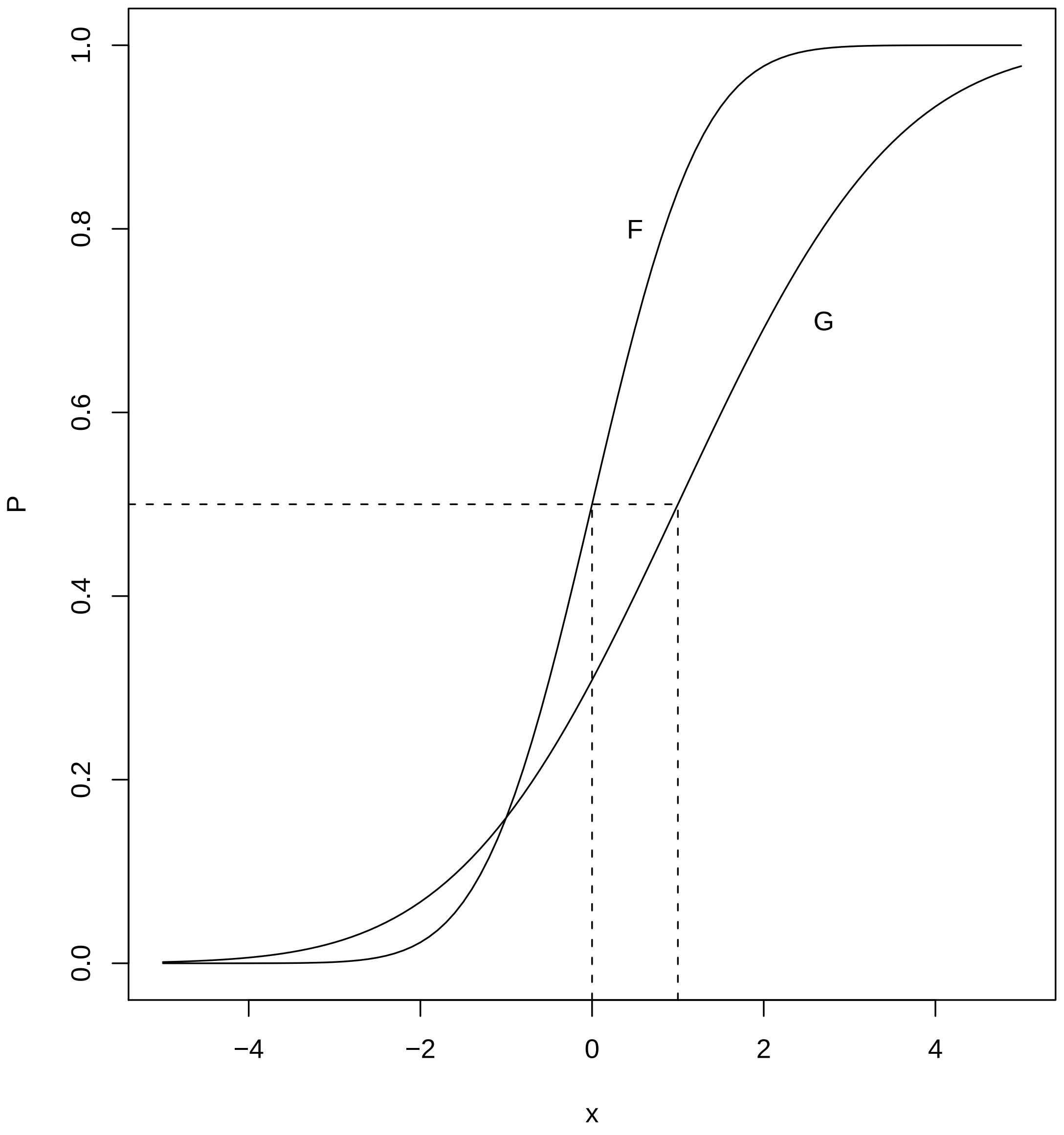
Then $\Delta(x)$ is uniquely defined as

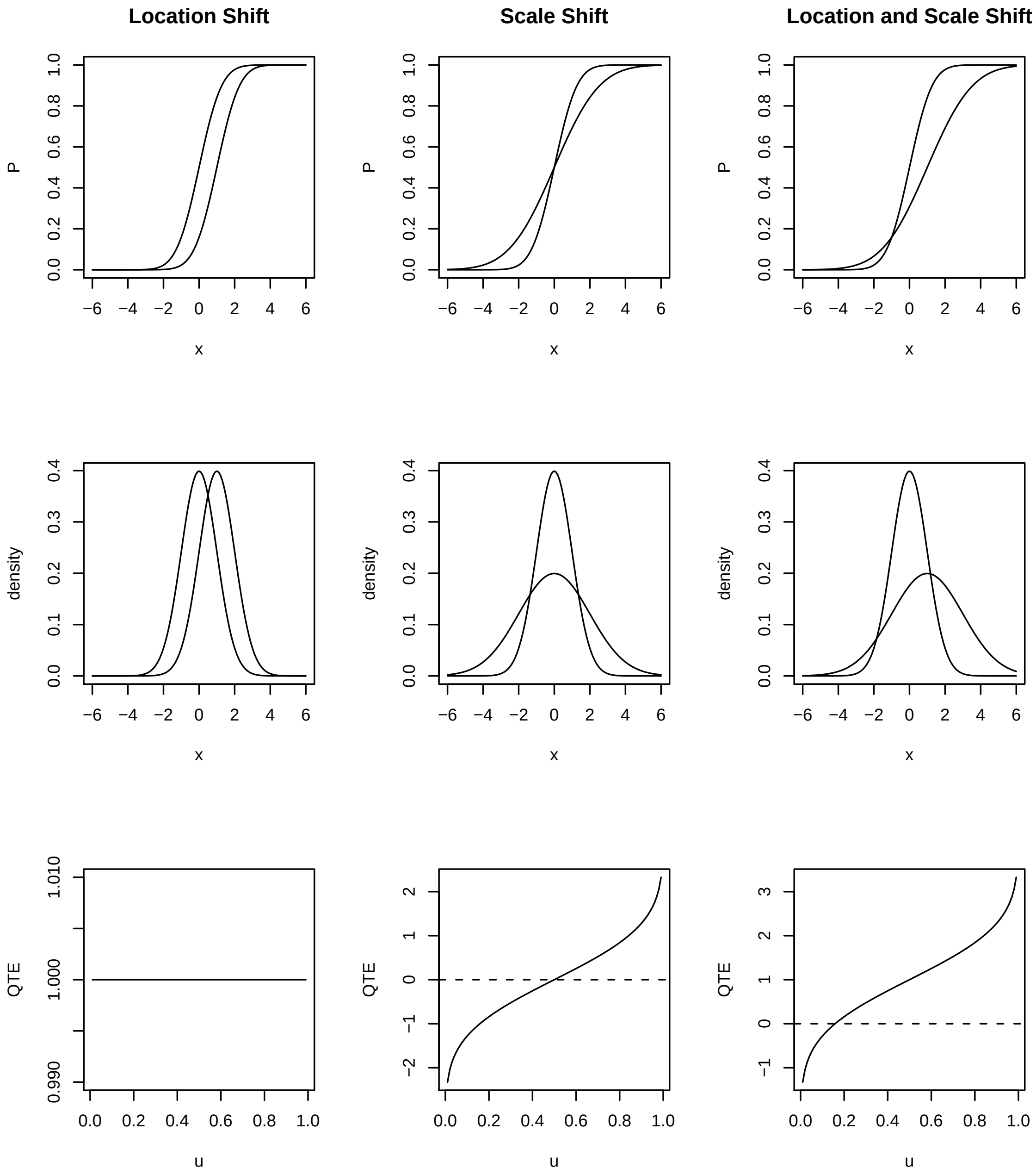
$$\Delta(x) = G^{-1}(F(x)) - x.$$

This is the essence of the conventional QQ-plot. Changing variables so $\tau = F(x)$ we have the quantile treatment effect (QTE):

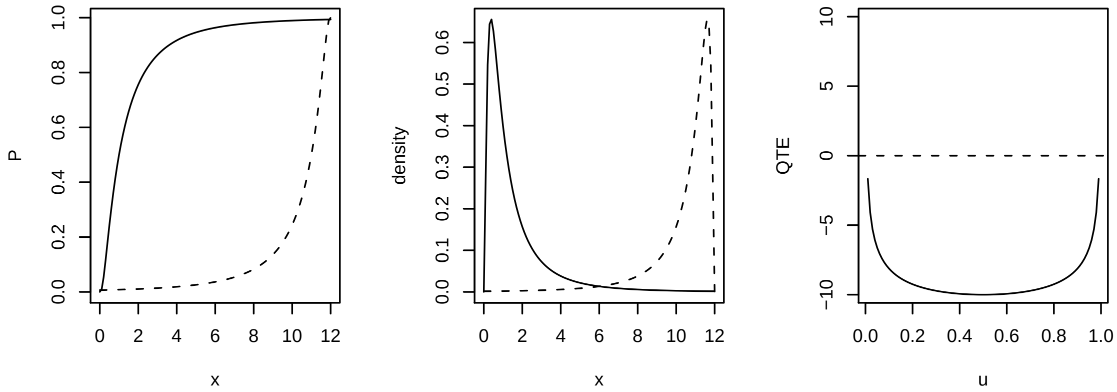
$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$

Lehmann-Doksum QTE


F
G
P
x
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4


Location Shift
Scale Shift
Location and Scale Shift
P
0.0
0.2
0.4
0.6
0.8
1.0
x
−6
−4
−2
0
2
4
6
density
0.0
0.1
0.2
0.3
0.4
QTE
0.990
1.000
1.010
−2
−1
0
1
2
3
u
0.0
0.2
0.4
0.6
0.8
1.0

An Asymmetric Example



Treatment shifts the distribution from right skewed to left skewed making the QTE U-shaped.

QTE via Quantile Regression

The Lehmann QTE is naturally estimable by

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau)$$

where $\hat{G}_n$ and $\hat{F}_m$ denote the empirical distribution functions of the treatment and control observations, Consider the quantile regression model

$$Q_{Y_i}(\tau | D_i) = \alpha(\tau) + \delta(\tau) D_i$$

where D_i denotes the treatment indicator, and $Y_i = h(T_i)$, *e.g.* $Y_i = \log T_i$, which can be estimated by solving,

$$\min \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \delta D_i)$$

Four Applications

- Engel's Law: A Classical Economic Example
- Infant Birthweight: A Public Health Example
- Melbourne Daily Temperature: A Time Series Example
- Infant and Adolescent Growth Charts

Engel's Food Expenditure Data

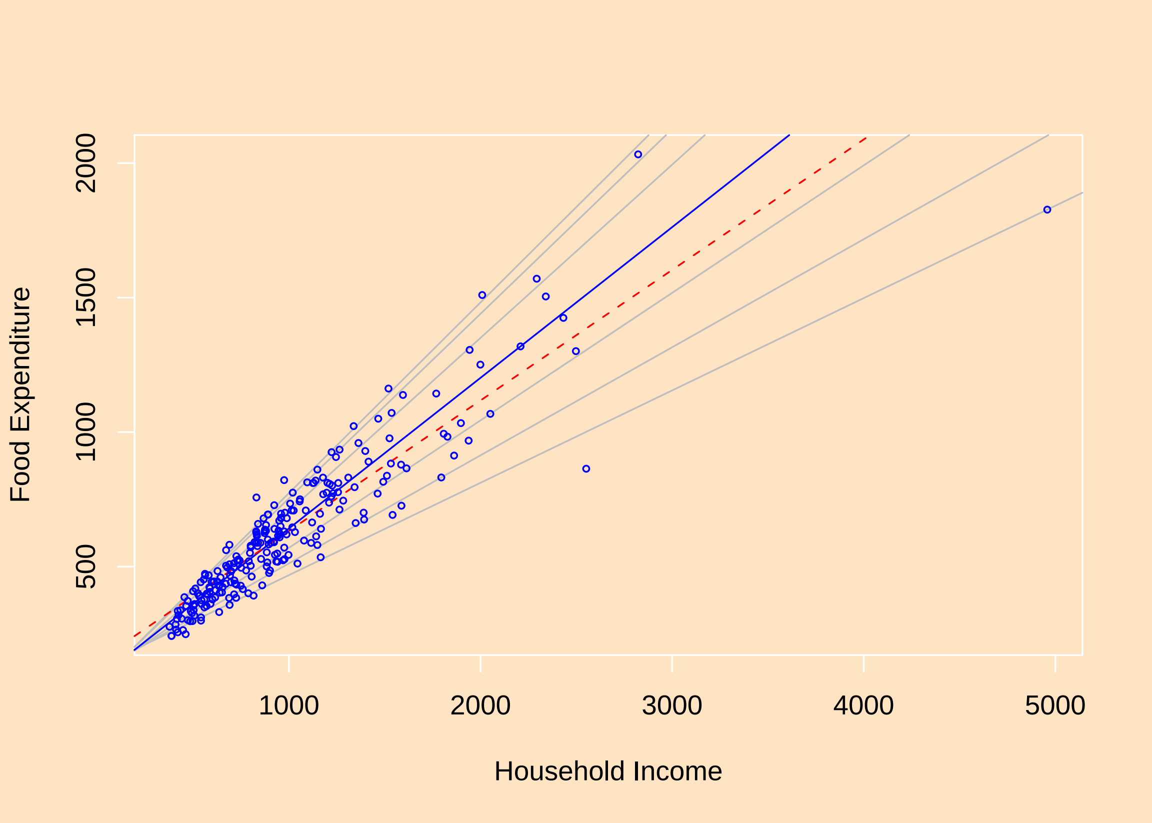


Figure 1: Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.

Engel's Food Expenditure Data

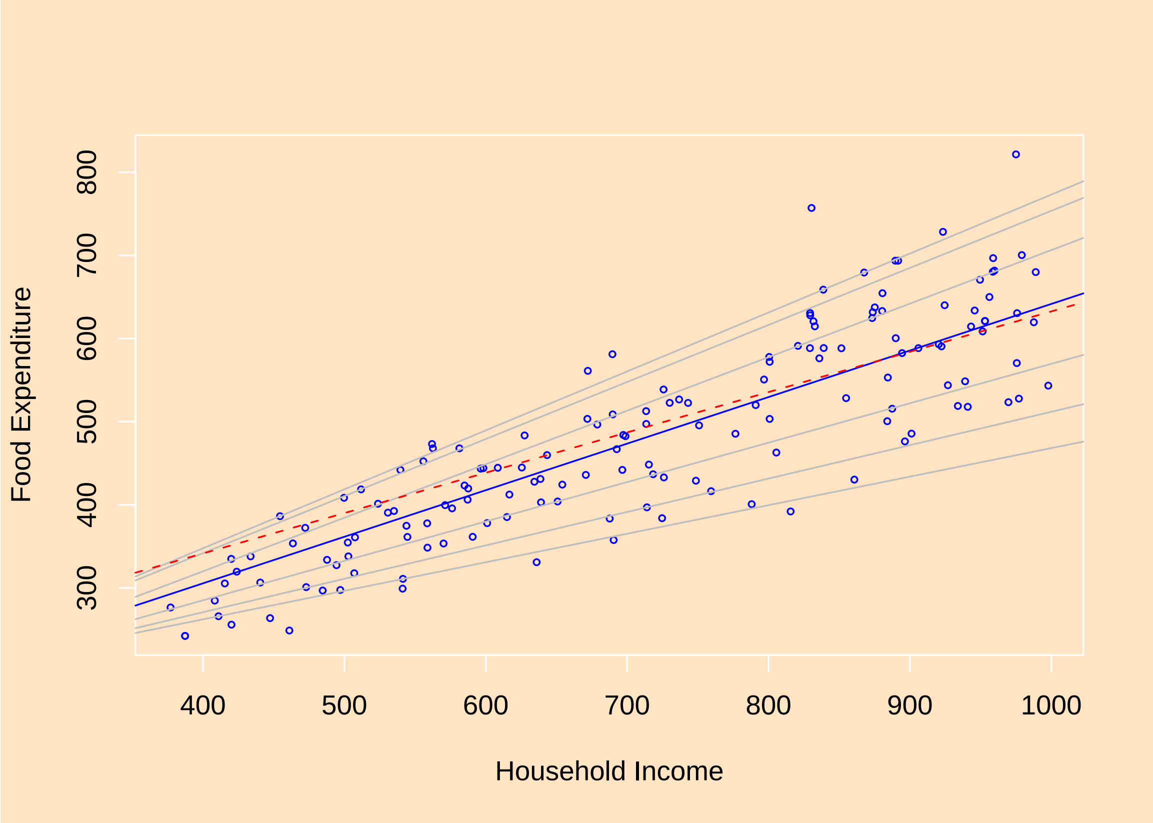
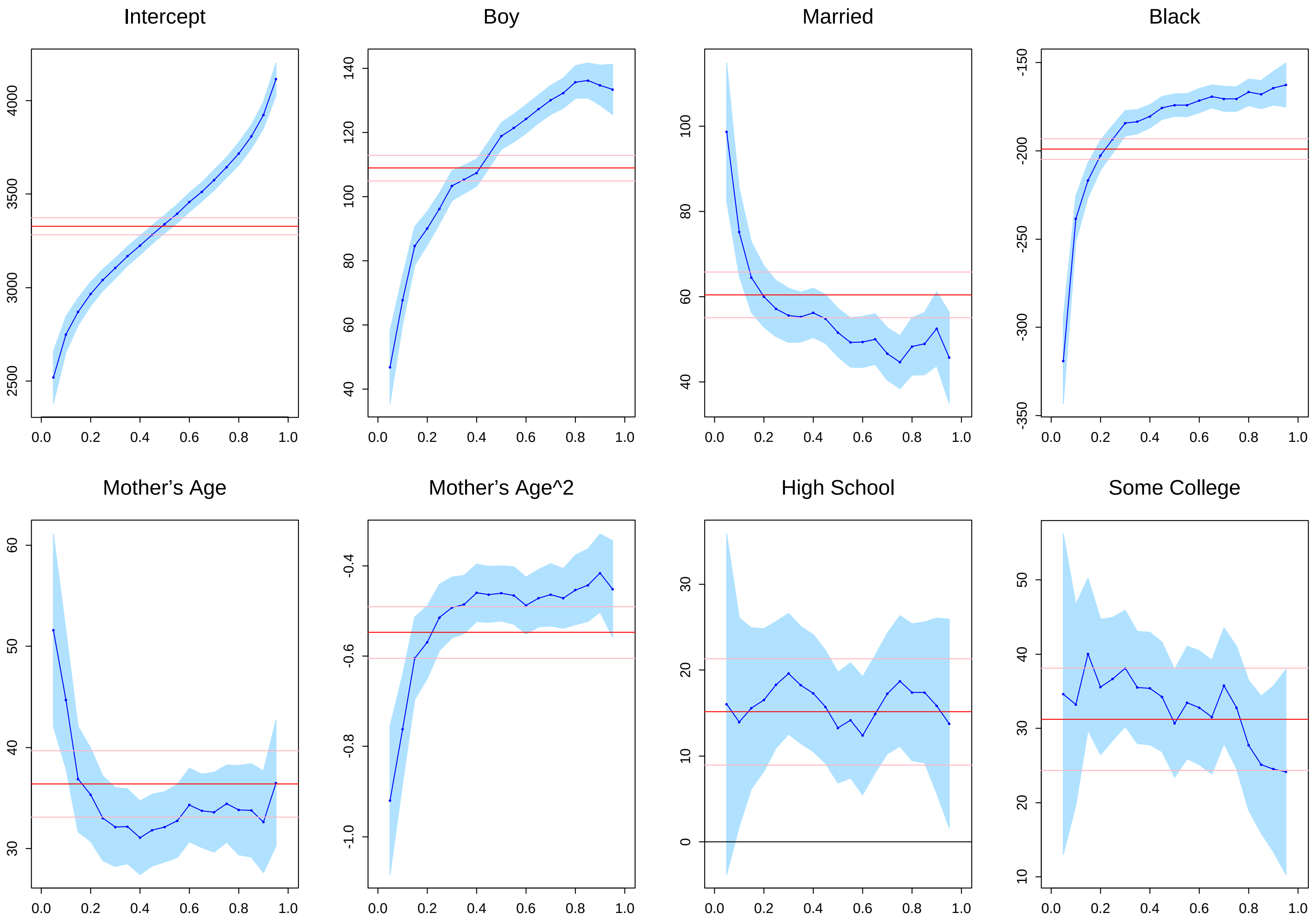


Figure 2: Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.

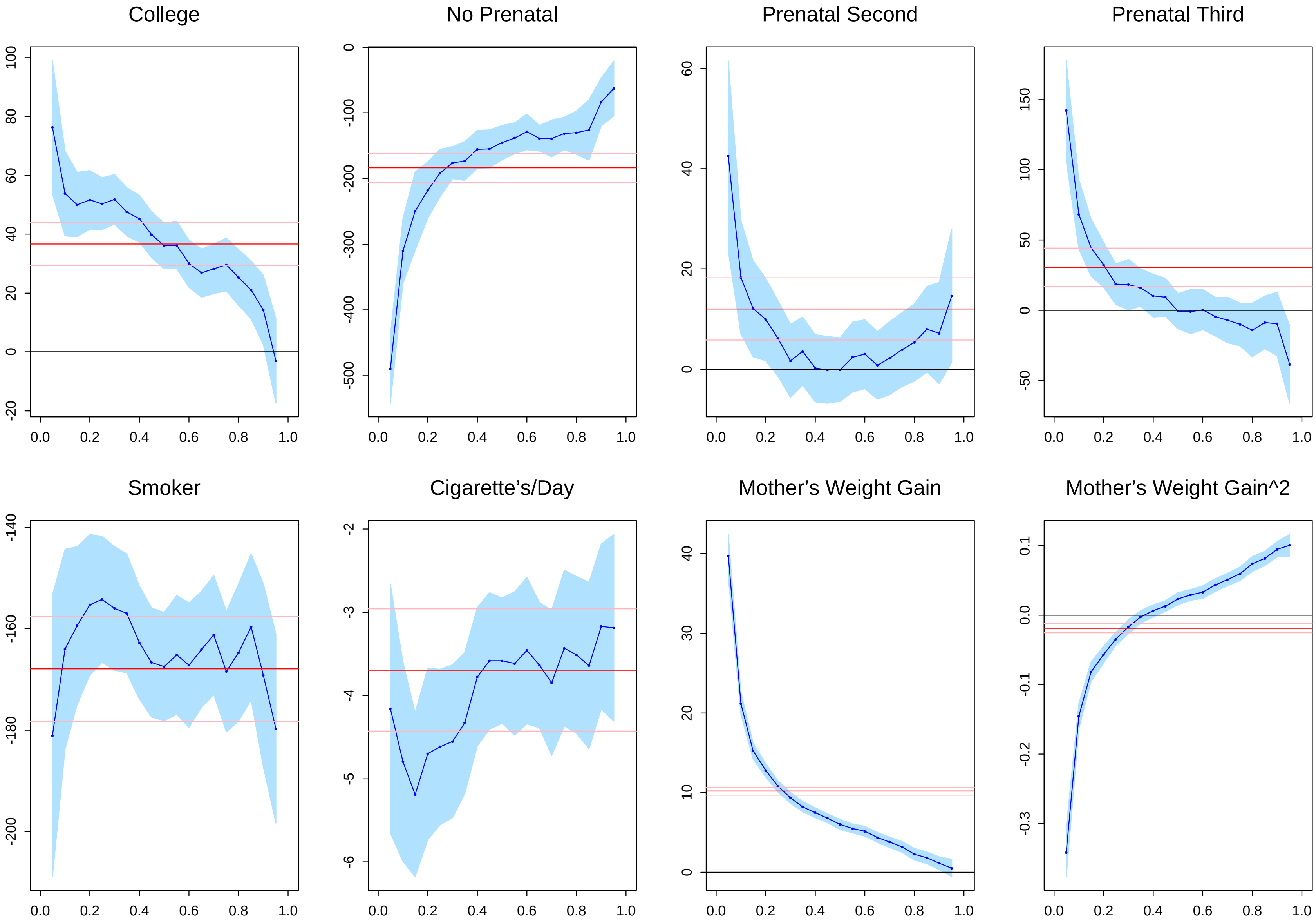
A Model of Infant Birthweight

- Reference: Abrevaya (2001), Koenker and Hallock (2001)
- Data: June, 1997, Detailed Natality Data of the US. Live, singleton births, with mothers recorded as either black or white, between 18-45, and residing in the U.S. Sample size: 198,377.
- Response: Infant Birthweight (in grams)
- Covariates:
  - ⋆ Mother's Education
  - ⋆ Mother's Prenatal Care
  - ⋆ Mother's Smoking
  - ⋆ Mother's Age
  - ⋆ Mother's Weight Gain

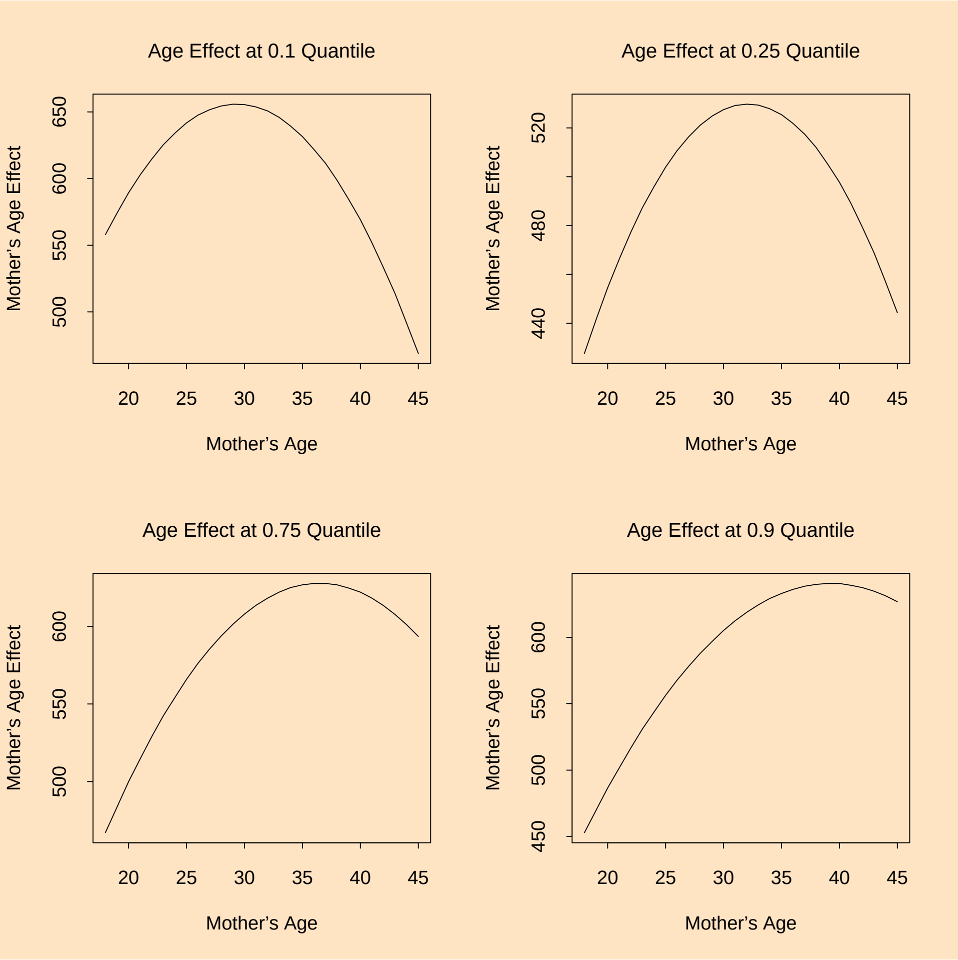
Quantile Regression Birthweight Model I



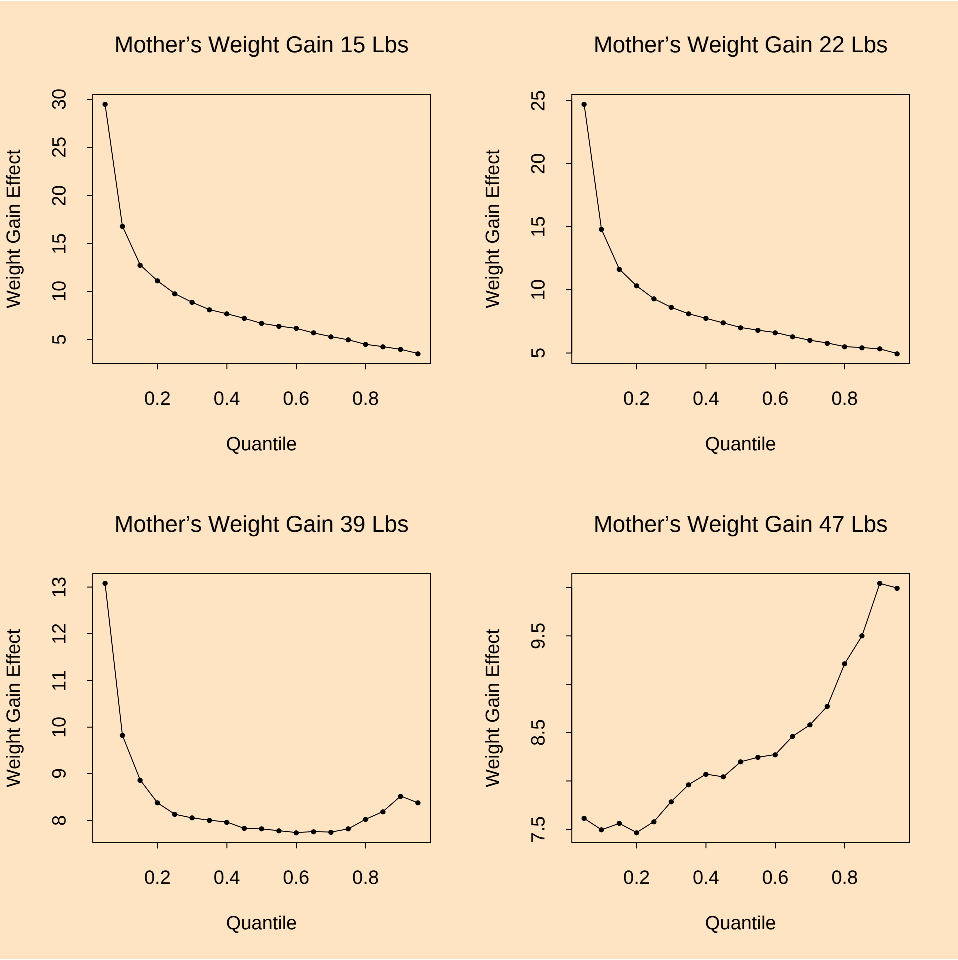
Quantile Regression Birthweight Model II



Marginal Effect of Mother's Age



Marginal Effect of Mother’s Weight Gain



AR(1) Model of Melbourne Daily Temperature

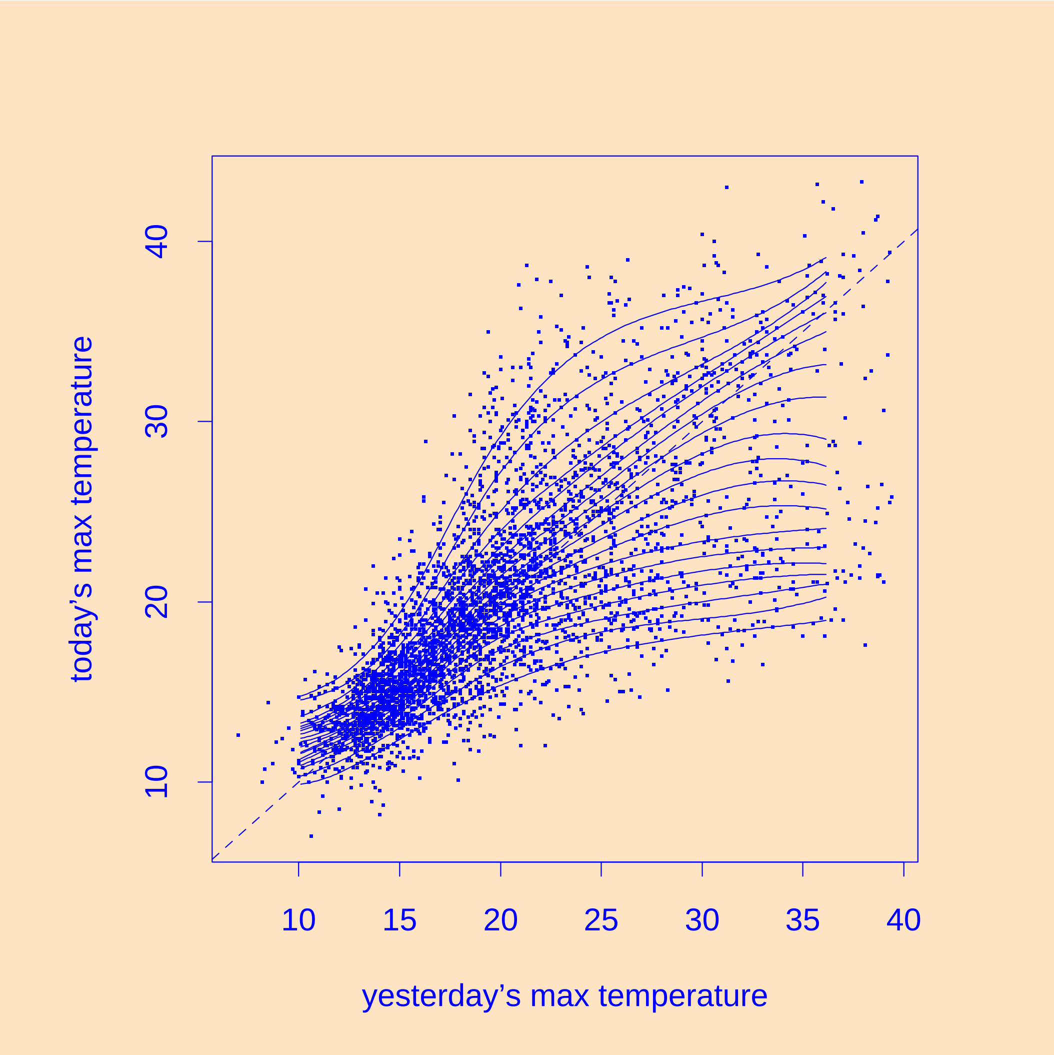
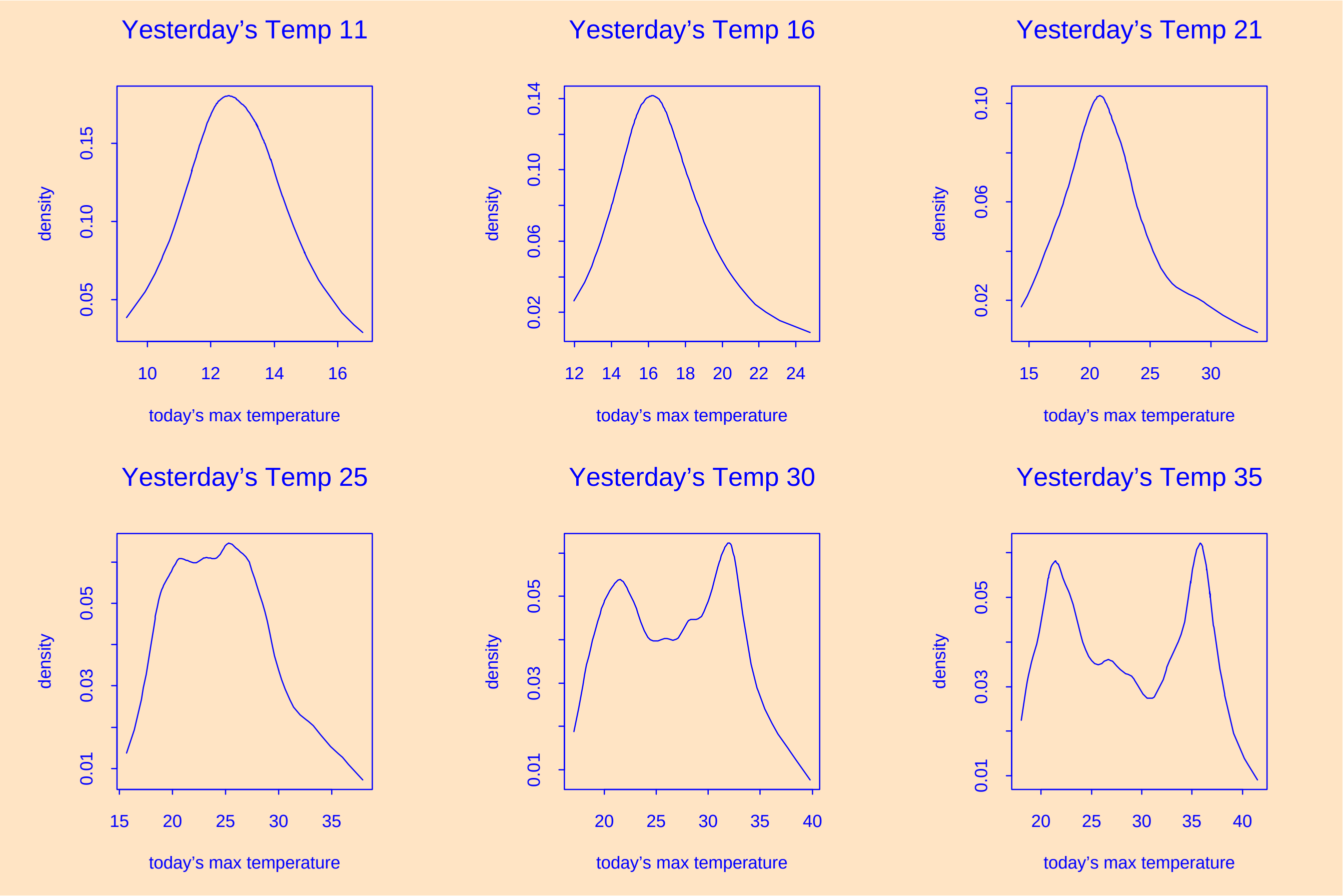
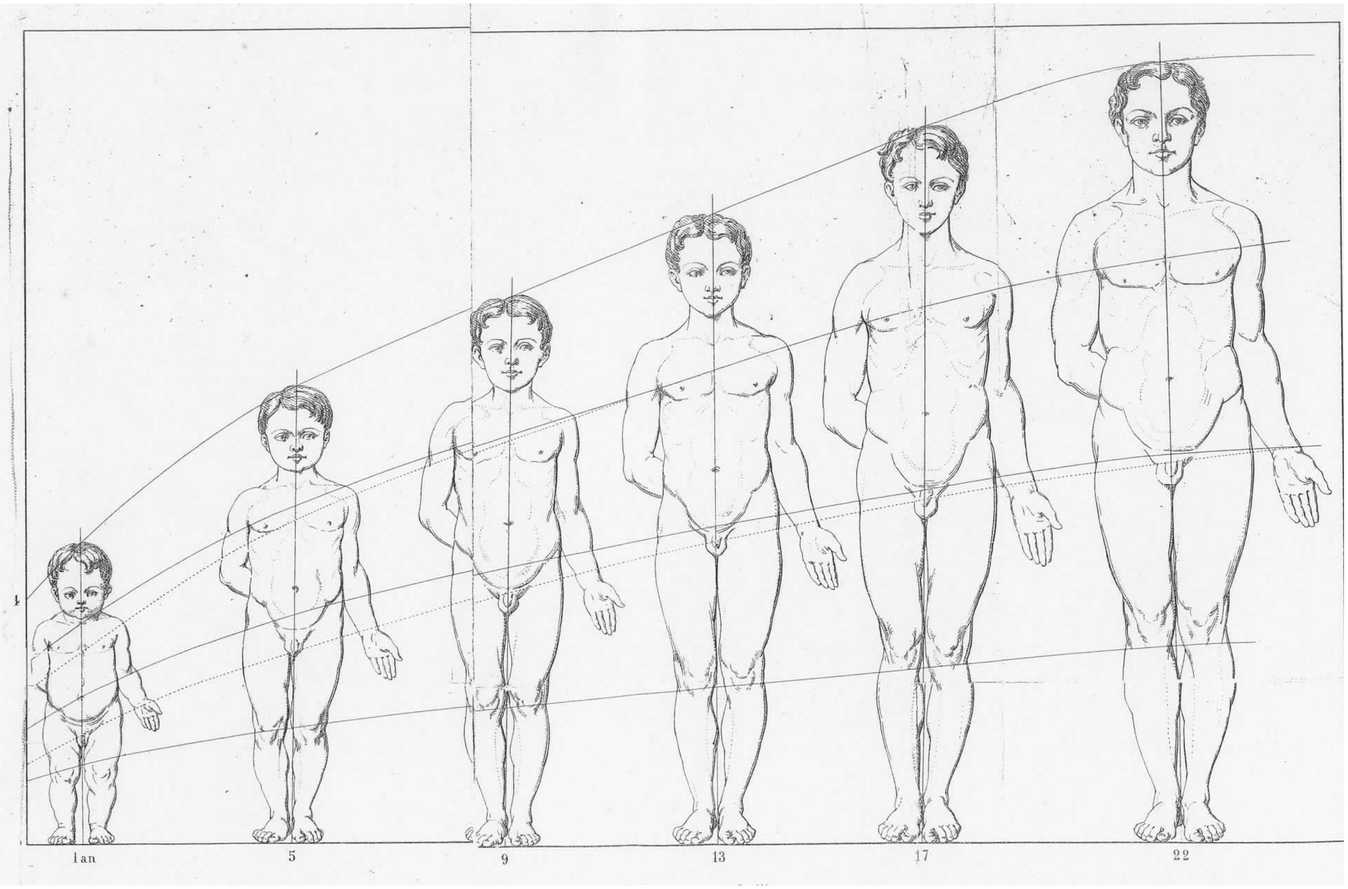


Figure 3: The plot illustrates 10 years of daily maximum temperature data for Melbourne, Australia as an AR(1) scatterplot. Superimposed are estimated conditional quantile functions for $\tau \in \{.05, .10, ..., .95\}$.

Conditional Densities of Melbourne Daily Temperature



Quetelet's (1871) Growth Chart



Penalized Maximum Likelihood Estimation

- References: Cole (1988), Cole and Green (1992), and Carey(2002)
- Data: $\{Y_i(t_{i,j}) : \ j = 1, \ldots, J_i, \ i = 1, \ldots, n.\}$
- 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

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

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

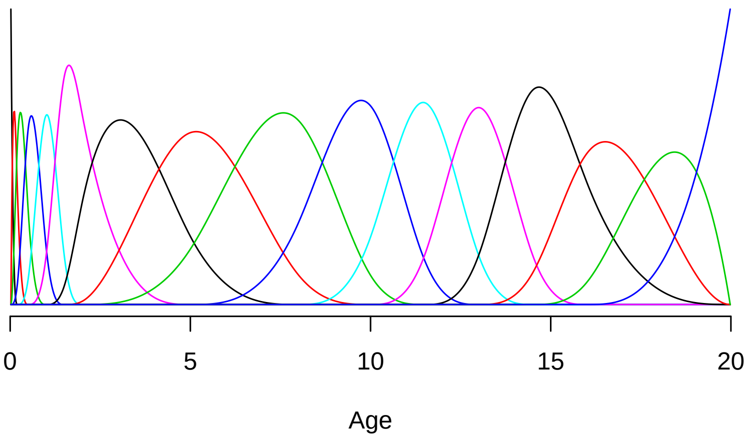
$$\hat{g}(\tau|x) = \text{argmin}_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i))$$

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:



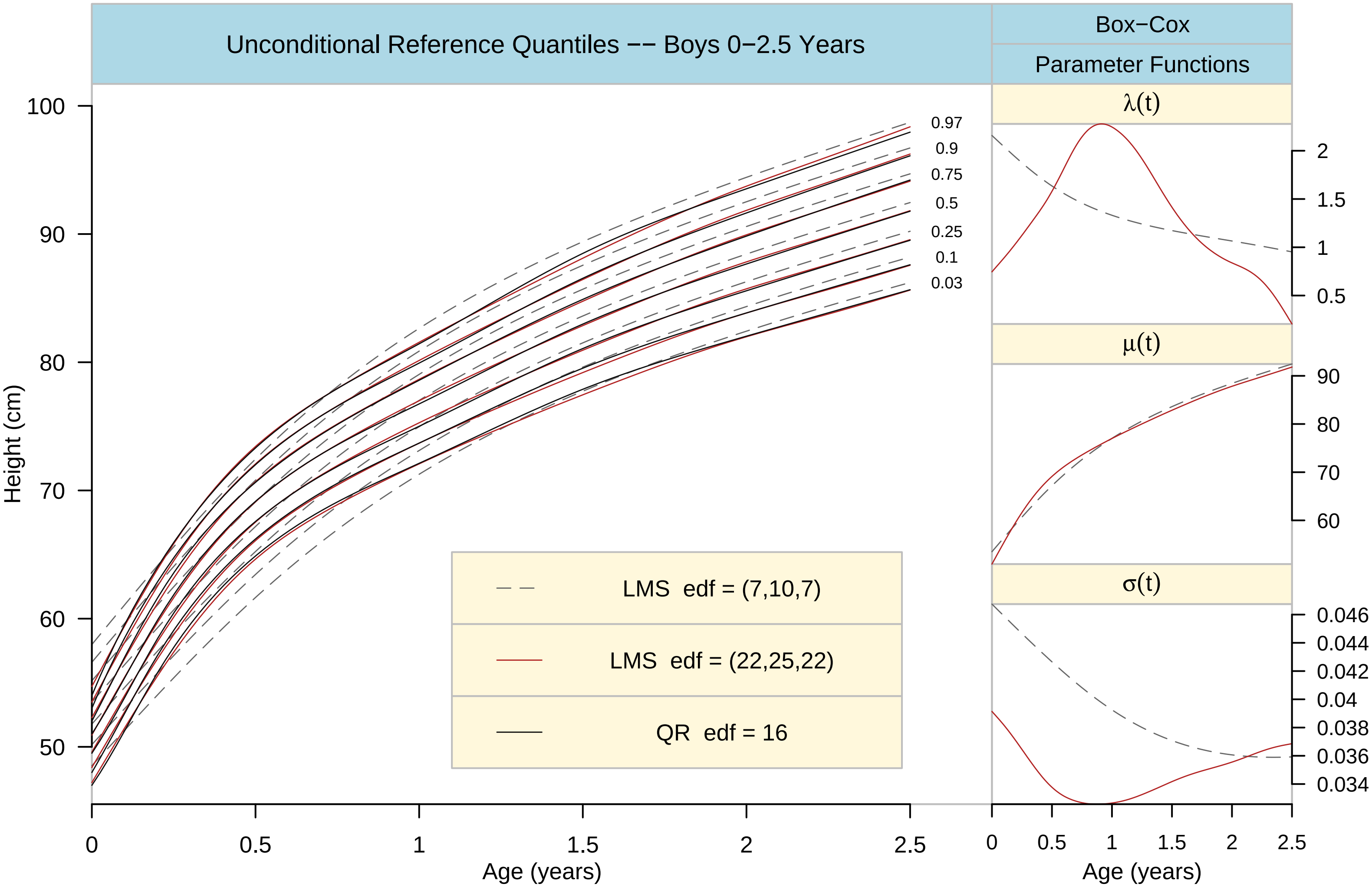
In R these quantile regression models can be estimated with the command.

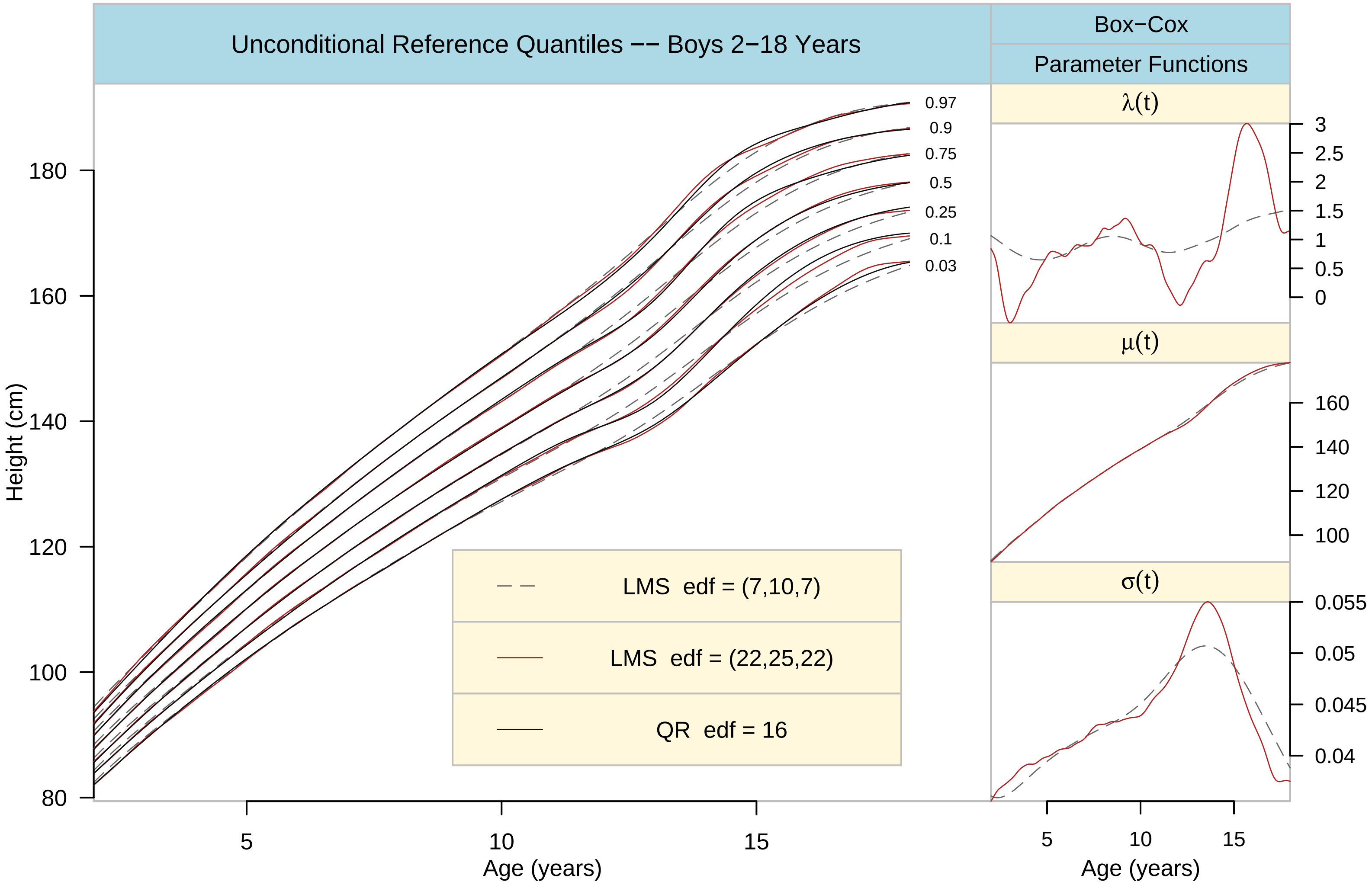
```
fit <- rq(y ∼ bs(x,knots=knots),tau = 1:9/10)
```

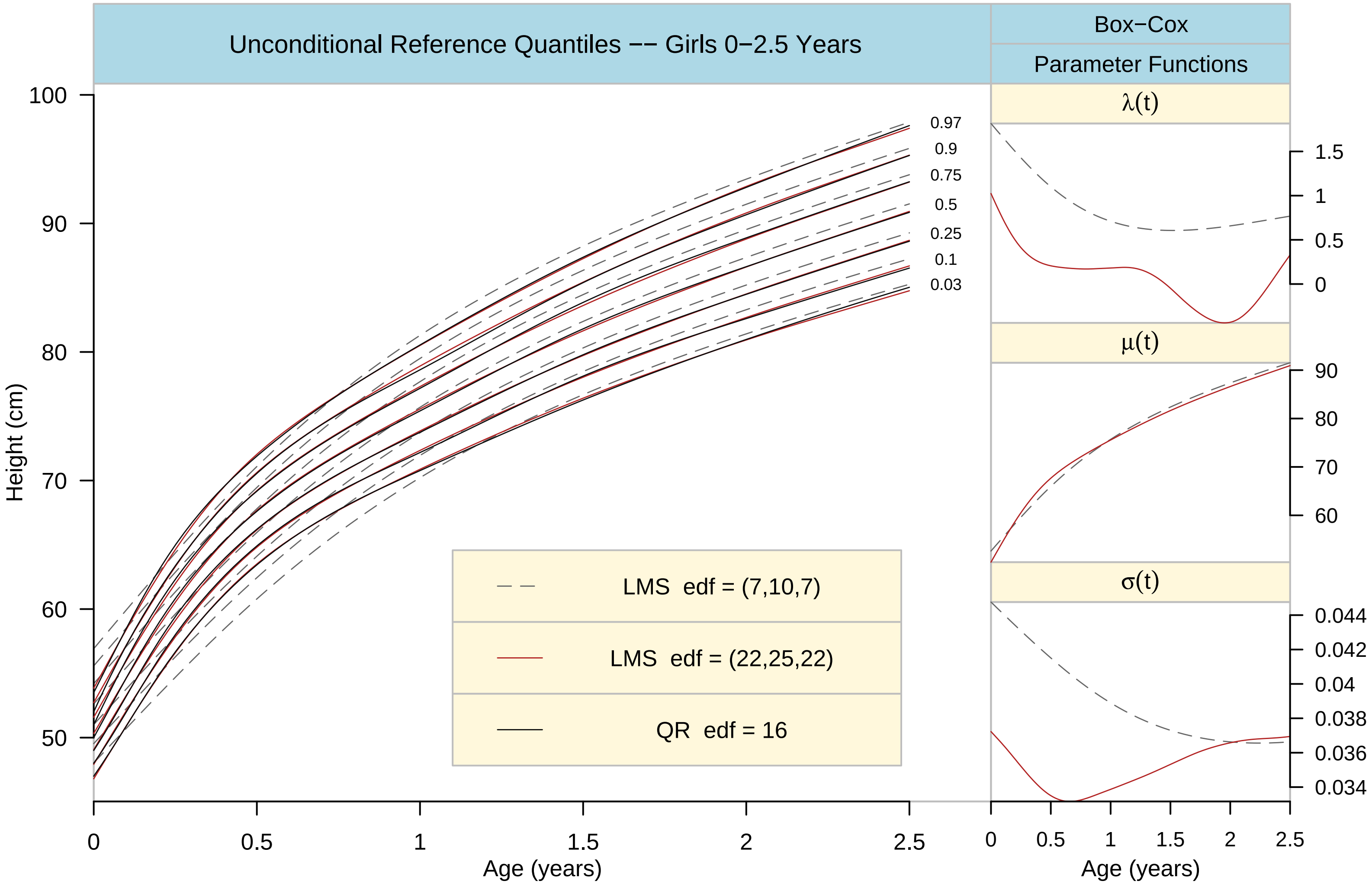
Similar functionality in SAS is coming “real soon now.”

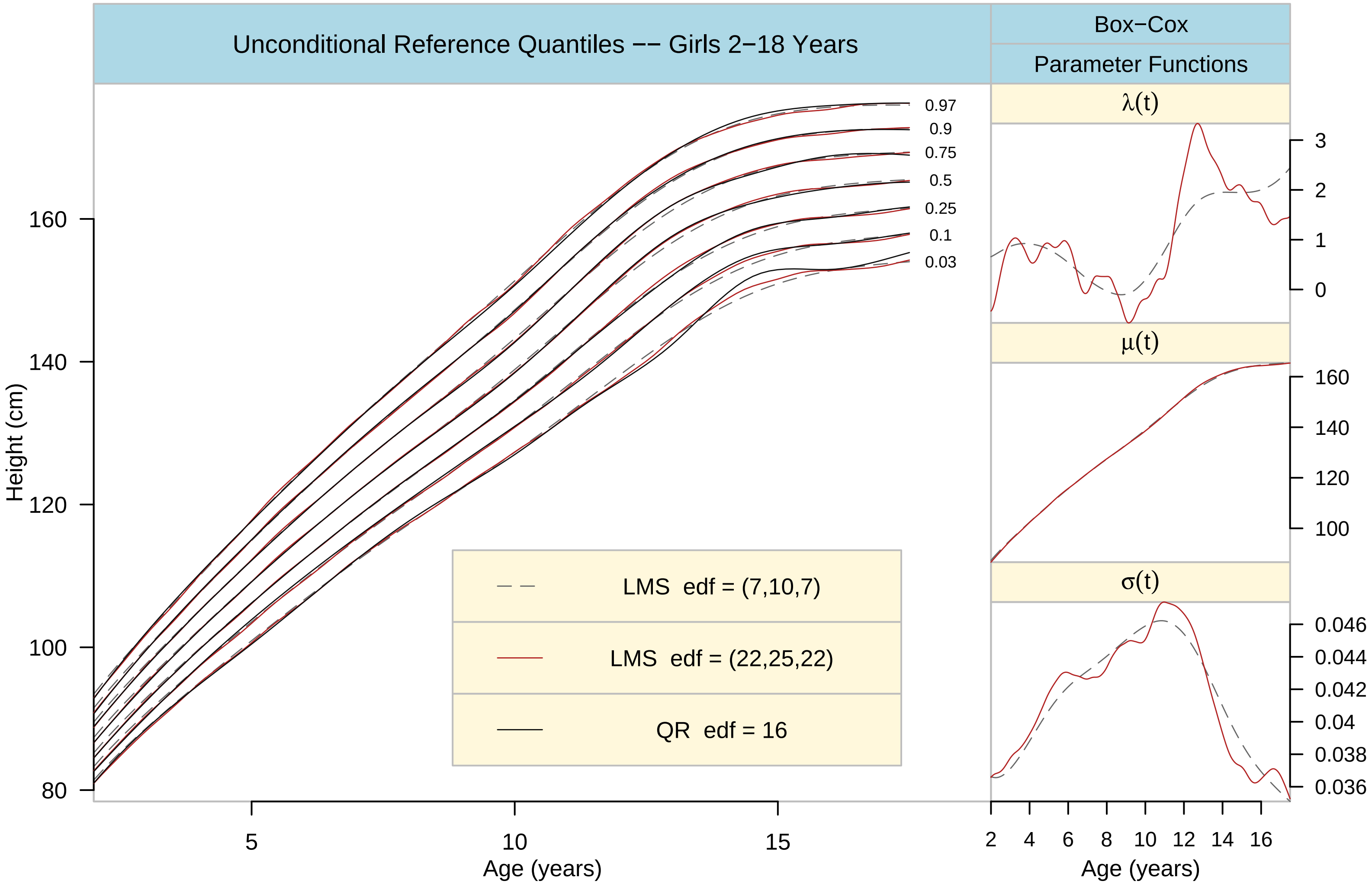
Data

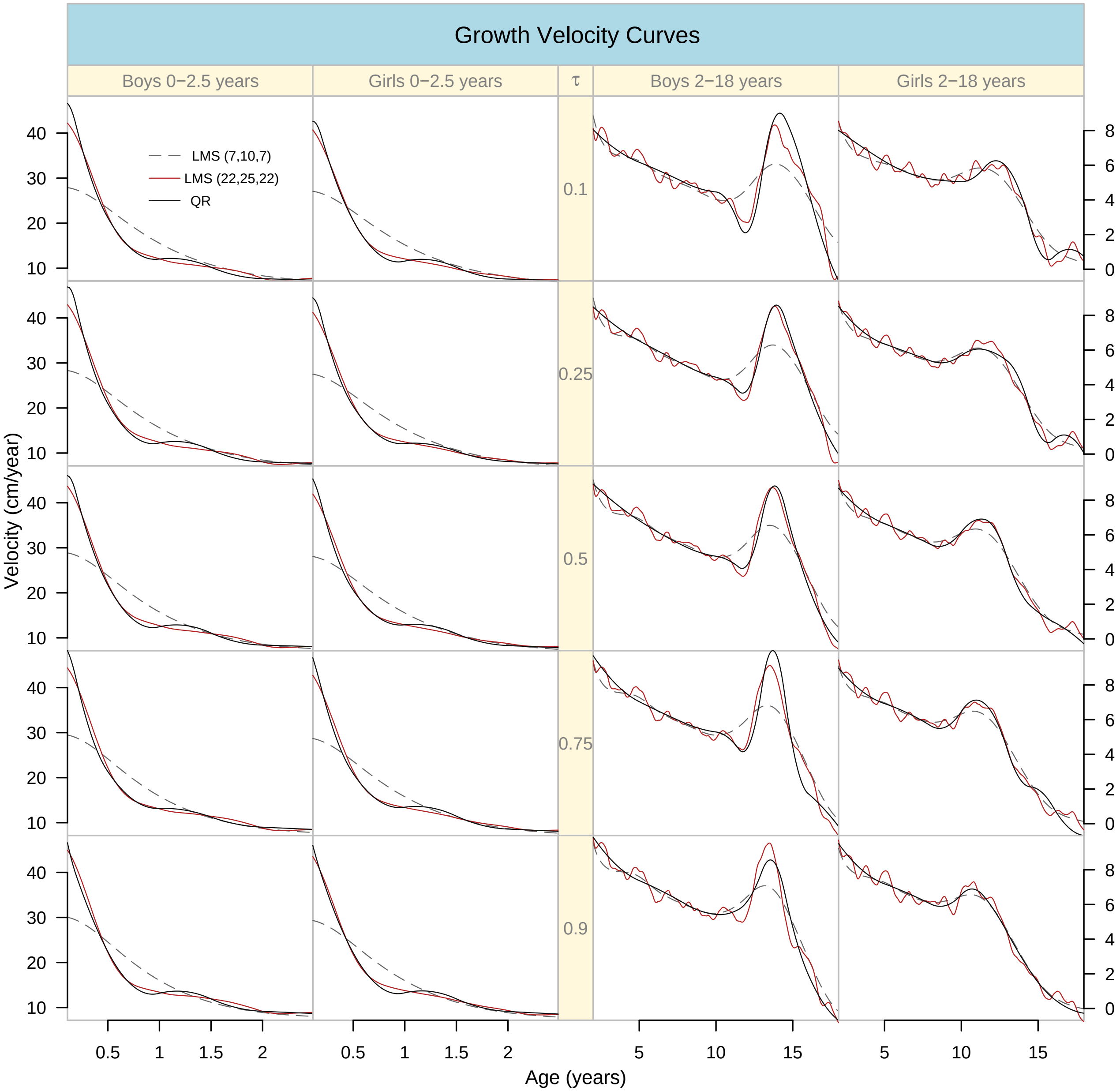
- Longitudinal measurements on height for 2514 Finnish children,
- 1143 boys, 1162 girls – all healthy, full-term, singleton births,
- About 20 measurements per child,
- Two cohorts: 1096 born between 1959-61, 1209 born between 1968-72
- Sample constitutes 0.5 percent of Finns born in these periods.

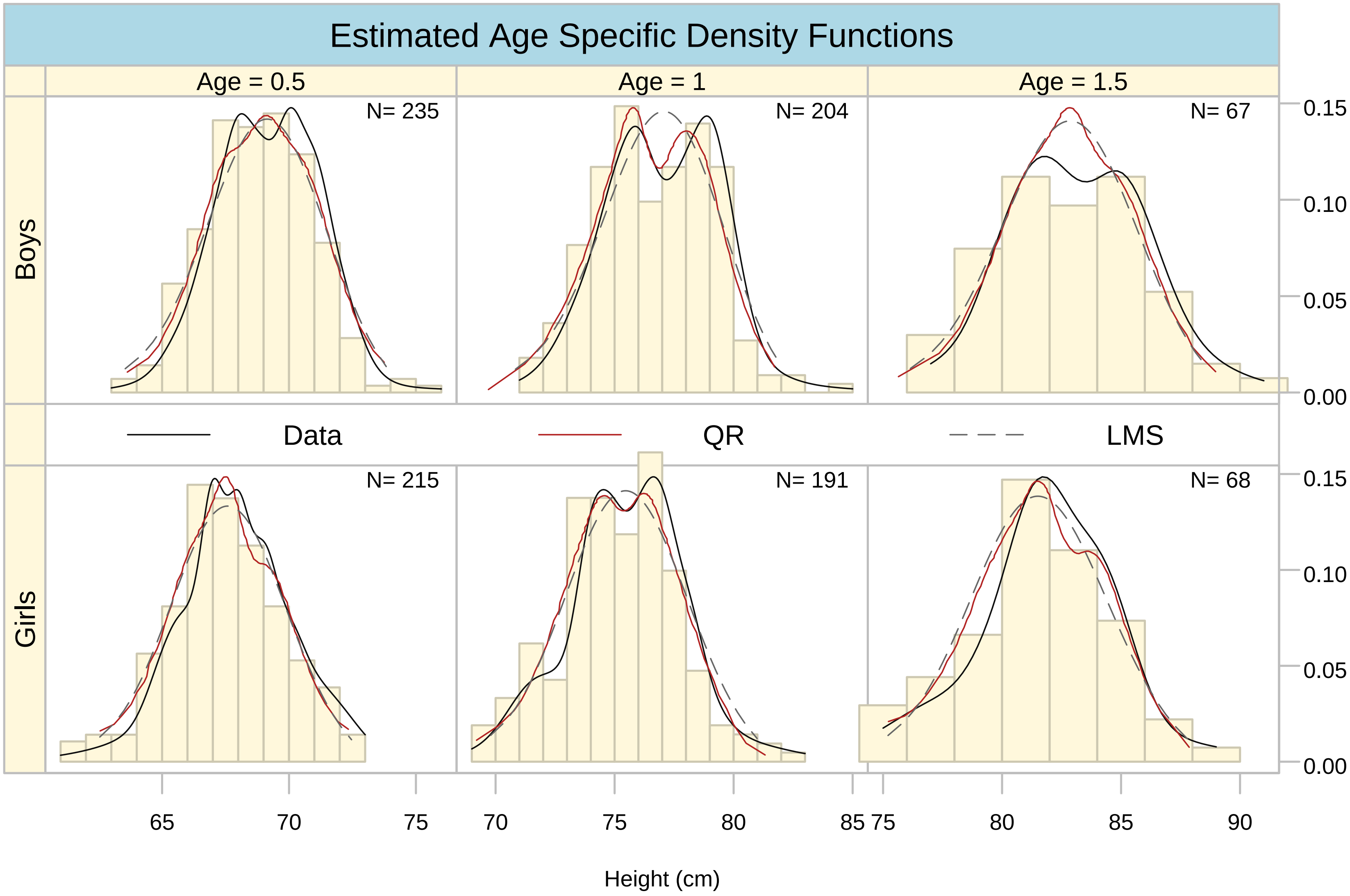

Unconditional Reference Quantiles –– Boys 0−2.5 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
100
90
80
70
60
50
0
0.5
1
1.5
2
2.5
Age (years)
0.97
0.9
0.75
0.5
0.25
0.1
0.03
LMS edf = (7,10,7)
LMS edf = (22,25,22)
QR edf = 16
2
1.5
1
0.5
90
80
70
60
0.046
0.044
0.042
0.04
0.038
0.036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

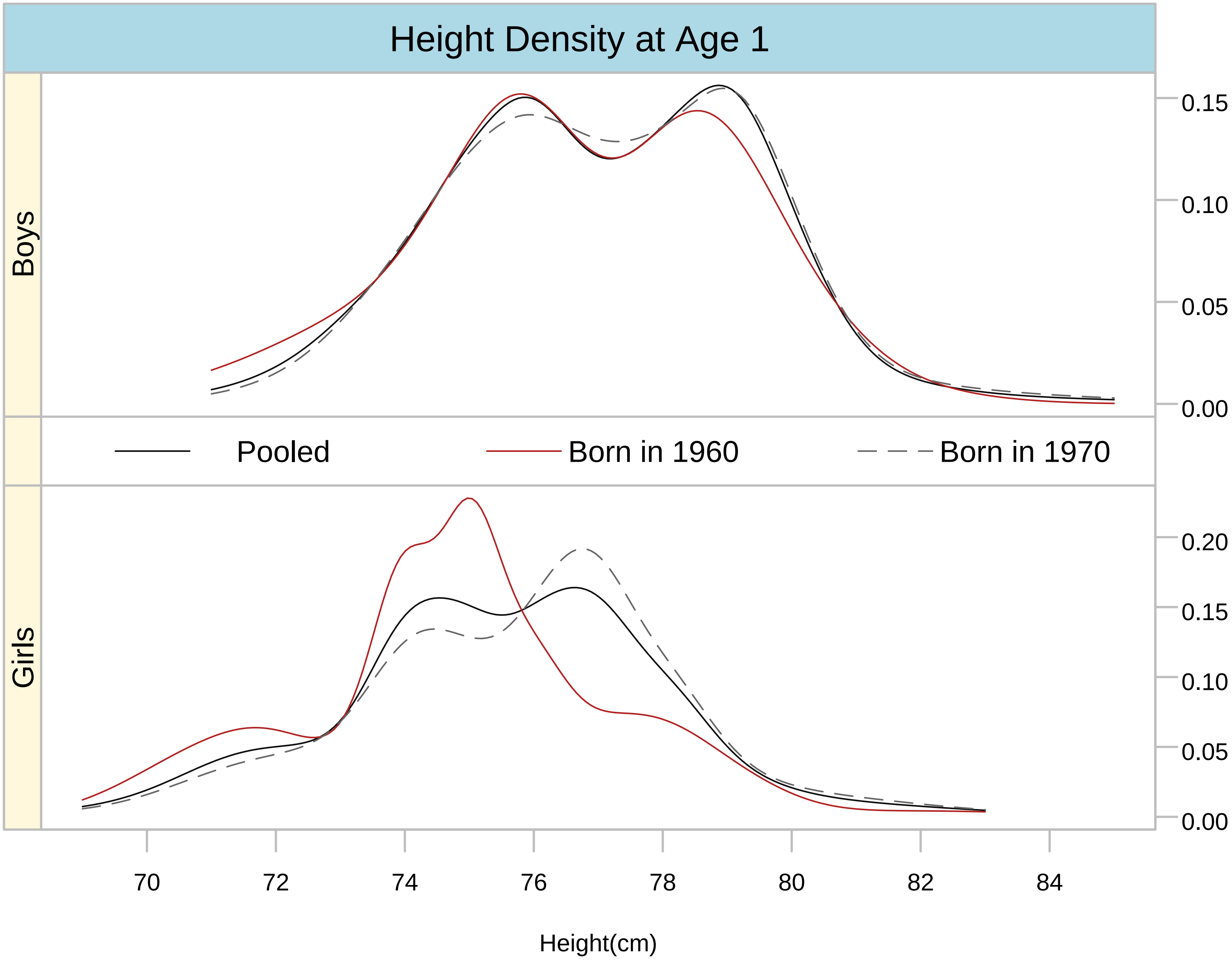

Unconditional Reference Quantiles −− Boys 2−18 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
80
100
120
140
160
180
5
10
15
Age (years)
0.97
0.9
0.75
0.5
0.25
0.1
0.03
LMS edf = (7,10,7)
LMS edf = (22,25,22)
QR edf = 16
3
2.5
2
1.5
1
0.5
0
160
140
120
100
0.055
0.05
0.045
0.04
5
10
15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


Unconditional Reference Quantiles –– Girls 0−2.5 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
100
90
80
70
60
50
0
0.5
1
1.5
2
2.5
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
1.5
1
0.5
0
90
80
70
60
0.044
0.042
0.04
0.038
0.036
0.034
0
0.5
1
1.5
2
2.5
Age (years)


Unconditional Reference Quantiles −− Girls 2−18 Years
Box−Cox
Parameter Functions
λ(t)
μ(t)
σ(t)
Height (cm)
80
100
120
140
160
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
1
0
160
140
120
100
0.046
0.044
0.042
0.04
0.038
0.036
2
4
6
8
10
12
14
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
10
20
30
40
0
2
4
6
8
0.5
1
1.5
2
5
10
15
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
0.15
0.10
0.05
0.00
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
0.15
0.10
0.05
0.00
0.20
0.15
0.10
0.05
0.00
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)

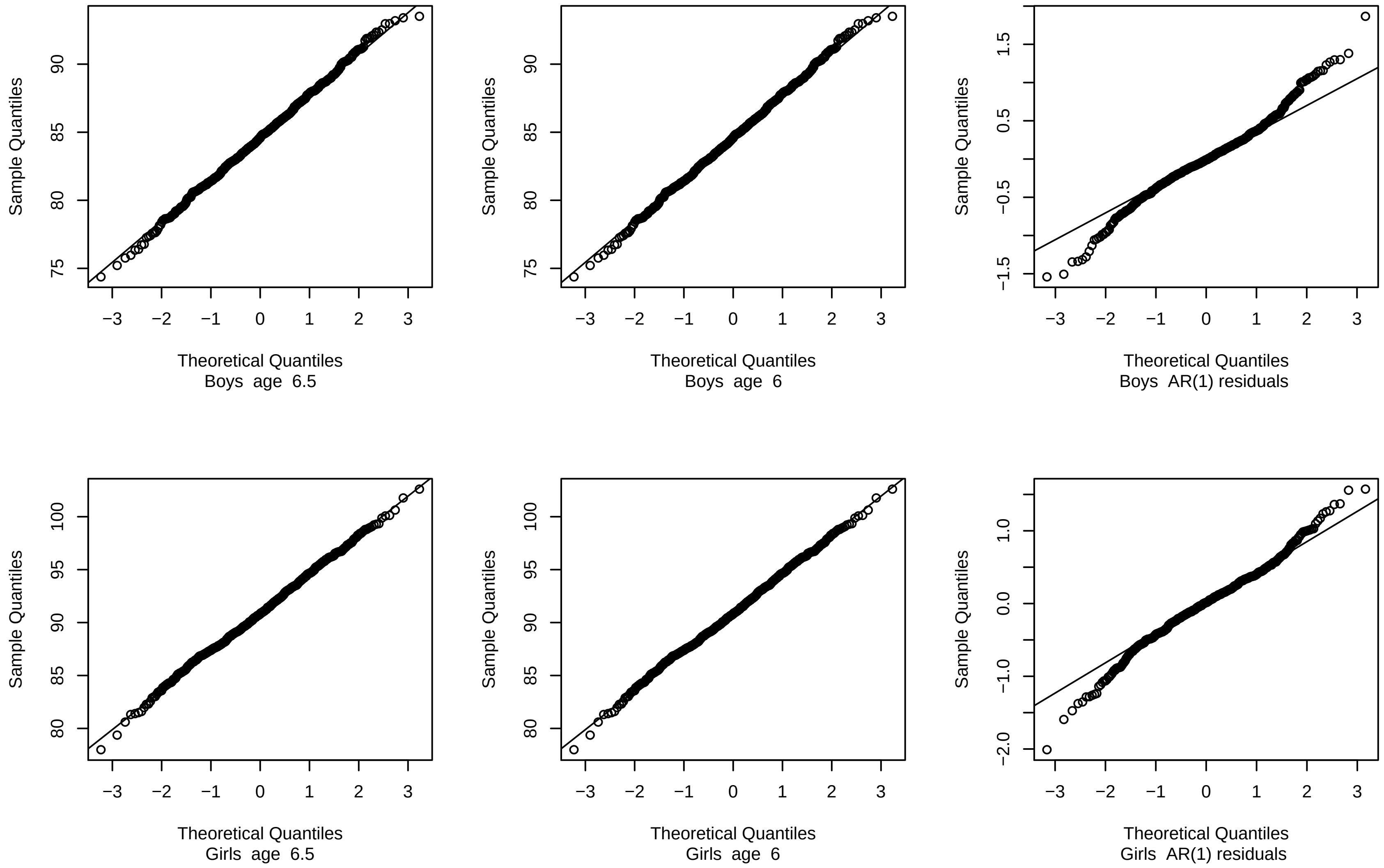
Transformation to Normality

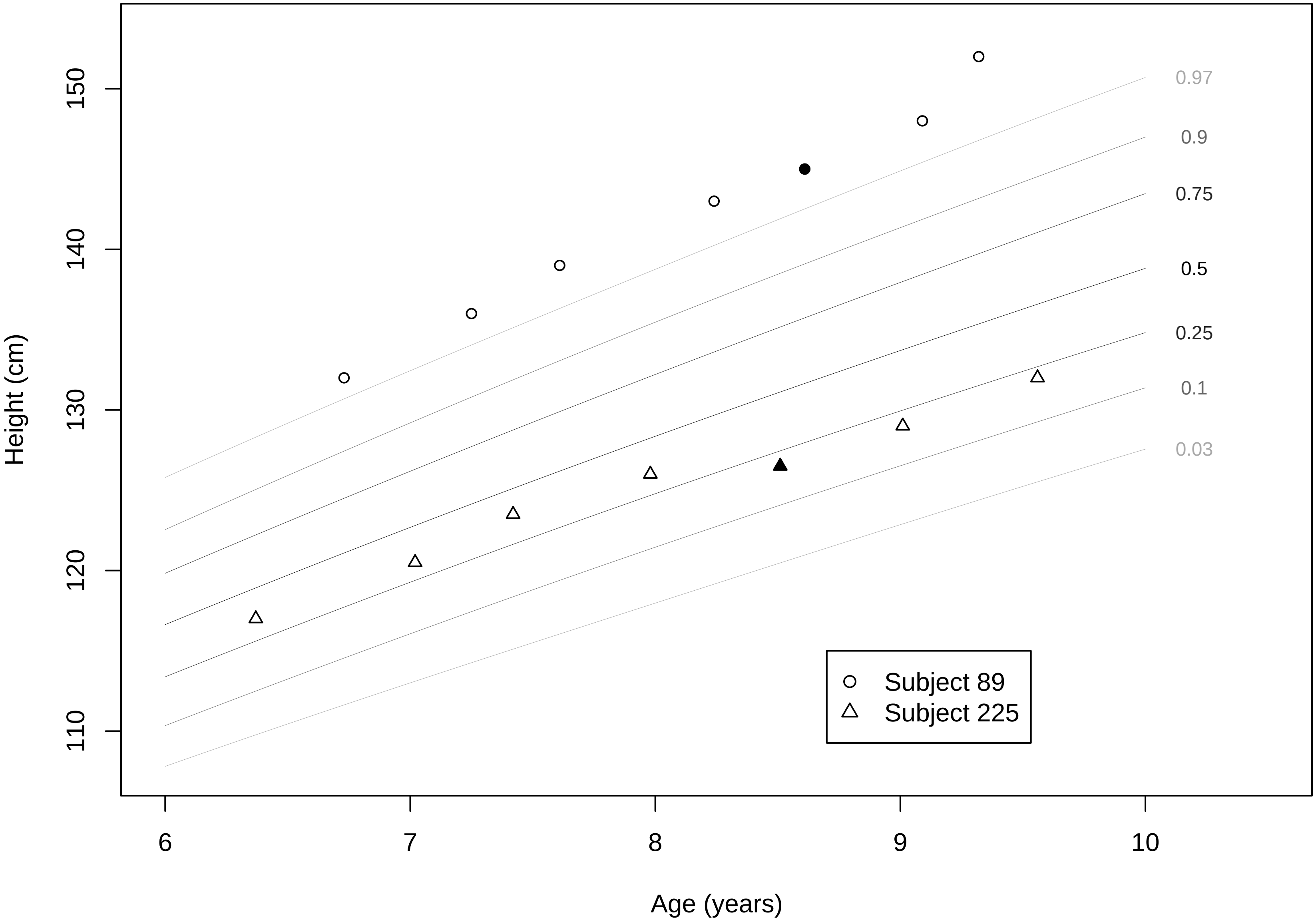
A presumed advantage of univariate (age-specific) transformation to normality is that once observations are transformed to univariate "Z-scores" they are automatically prepared to longitudinal autoregression:

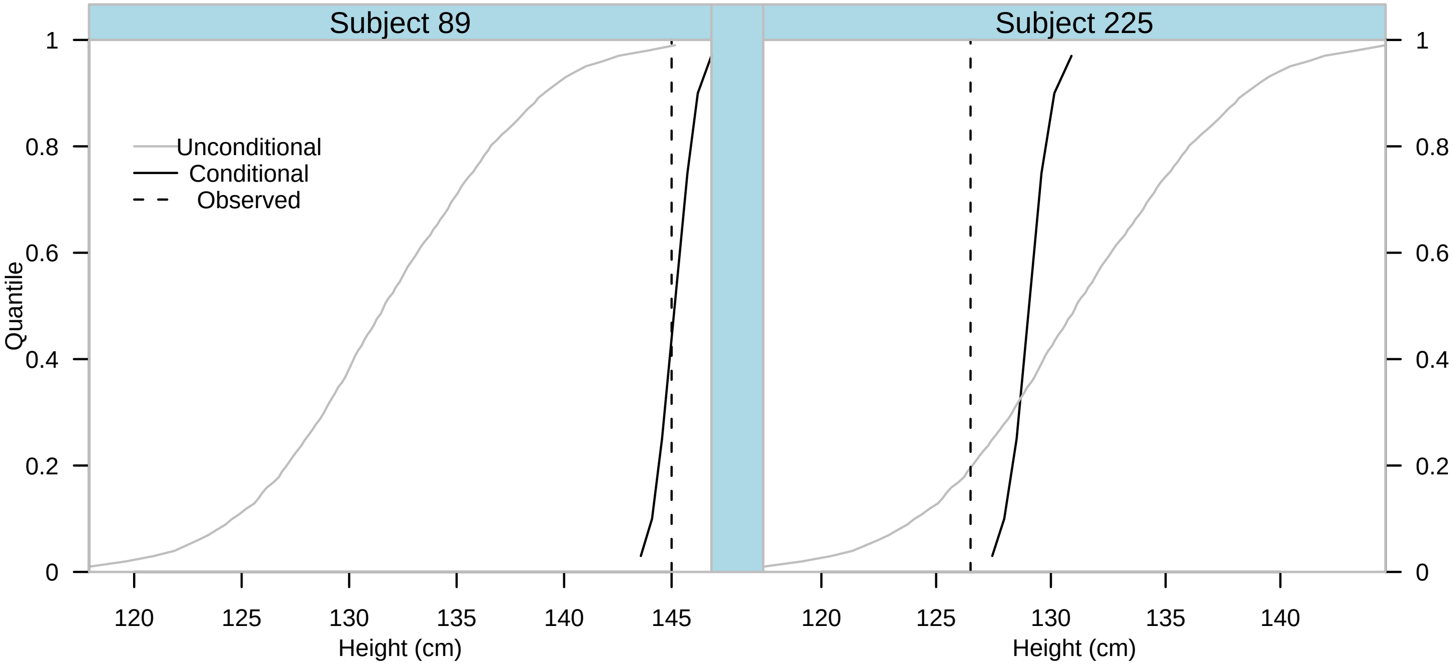
$$Z_t = \alpha_0 + \alpha_1 Z_{t-1} + U_t$$

Premise: Marginal Normality ⇒ 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
75
80
85
90
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Boys age 6.5
Sample Quantiles
75
80
85
90
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Boys age 6
Sample Quantiles
−1.5
−0.5
0.5
1.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Boys AR(1) residuals
Sample Quantiles
80
85
90
95
100
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Girls age 6.5
Sample Quantiles
80
85
90
95
100
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Girls age 6
Sample Quantiles
−2.0
−1.0
0.0
1.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Girls AR(1) residuals


Height (cm)
150
140
130
120
110
6
7
8
9
10
Age (years)
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
0
0.2
0.4
0.6
0.8
1
120
125
130
135
140
145
Height (cm)
120
125
130
135
140
Height (cm)

Quantile Regression for Growth Charts

- 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.

General Conclusions

- Quantile regression methods complement established mean regression (least-squares) methods.

- By focusing on local slices of the conditional distribution, they offer a useful deconstruction of conditional mean models.

- They provide a more flexible role for covariate effects allowing them to influence location, scale *and shape* of the response distribution.

- In applications a variety of unobserved heterogeneity phenomena are rendered observable.