Summer Course

Introduction to Quantile Regression

Roger Koenker
University of Illinois at Urbana-Champaign

La Roche-en-Ardennes: September, 2005

My thanks goes to many collaborators: Gib Bassett, Steve Portnoy, Pin Ng, Jana Jurečková , Jose Machado, Zhijie Xiao, Lingjie Ma, Gregory Kordas and Ivan Mizera, to the US National Science Foundation for continuing support, and to the open source software community for making the tools of my trade. The projected version of these slides were produced by ppower4 and pdflatex; all computations and graphics were done in the R language.

Motivation

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set.

Motivation

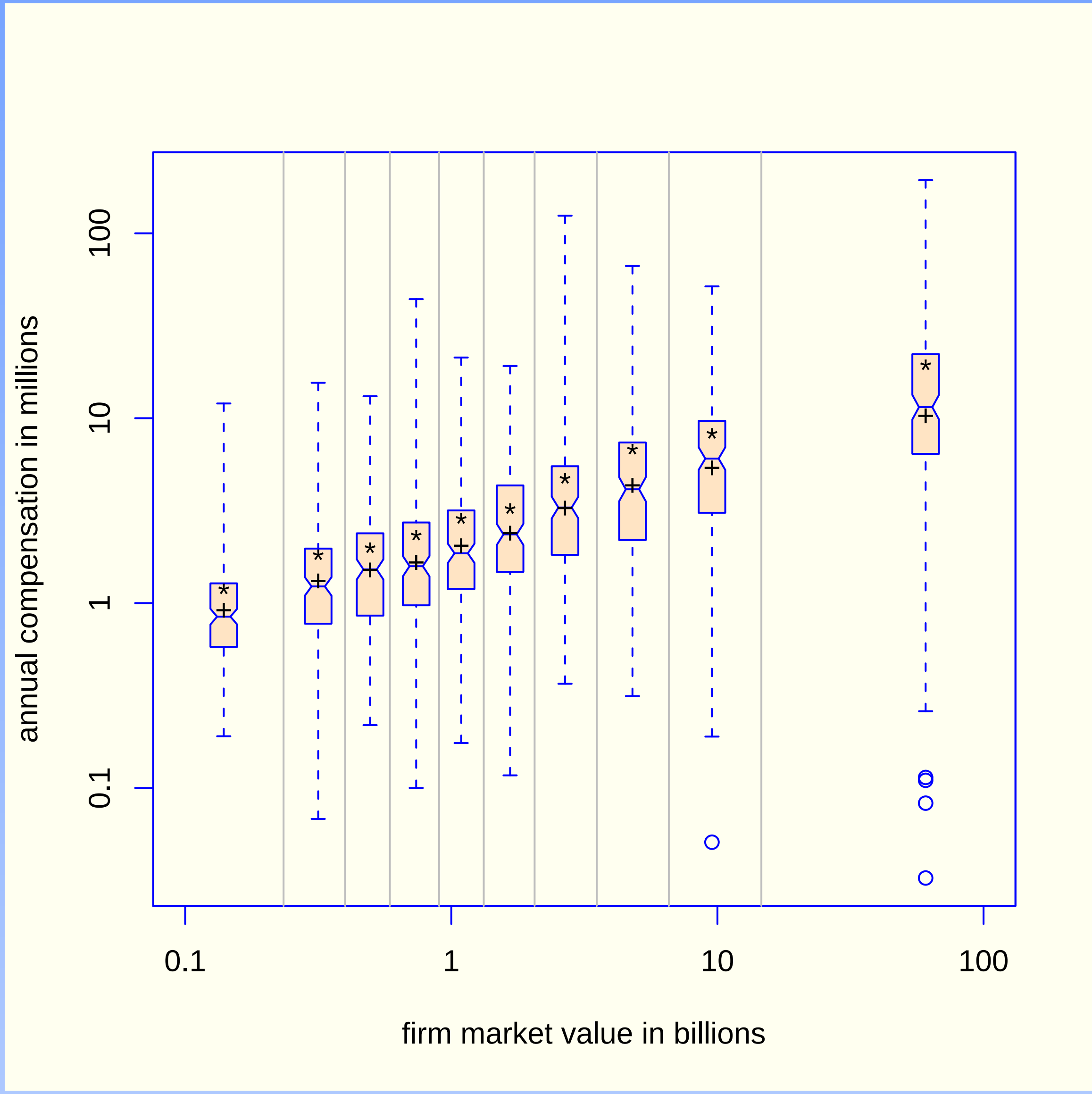
What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture.

Motivation

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

Mosteller and Tukey (1977)

Boxplot of CEO Pay by Firm Size



An Outline

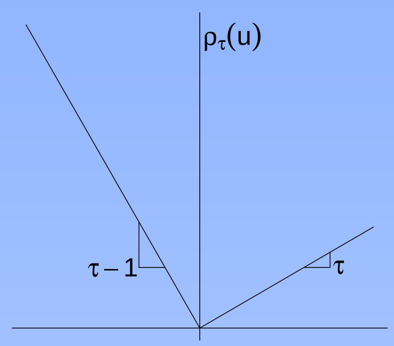
- An Historical Introduction to Regression
- What is Quantile Regression?
- Beyond Average Treatment Effects
- Two Artificial Examples
- Three Introductory Empirical Examples
  - ⋆ The Classical Engel Curve
  - ⋆ A Model of Infant Birthweight
  - ⋆ Maximum Daily Temperature in Melbourne

Sample Quantiles via Optimization

The τth sample quantile can be defined as any solution to:

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

where $\rho_\tau(u) = (\tau - I(u < 0))u$ as illustrated below.



The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_{m} E(Y - m)^2$$

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

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

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

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

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

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

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.*

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

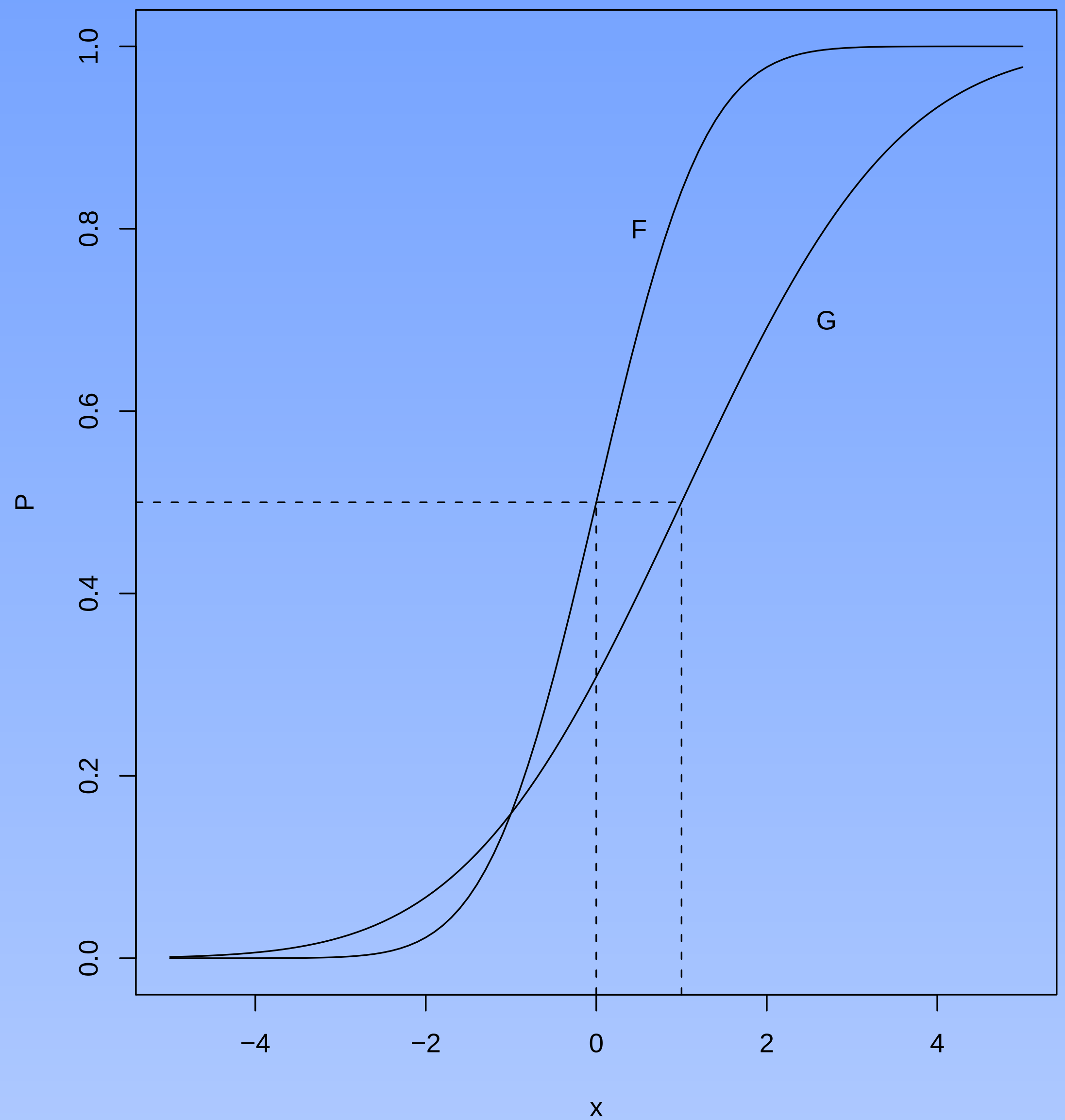
Then $\Delta(x)$ is uniquely defined as

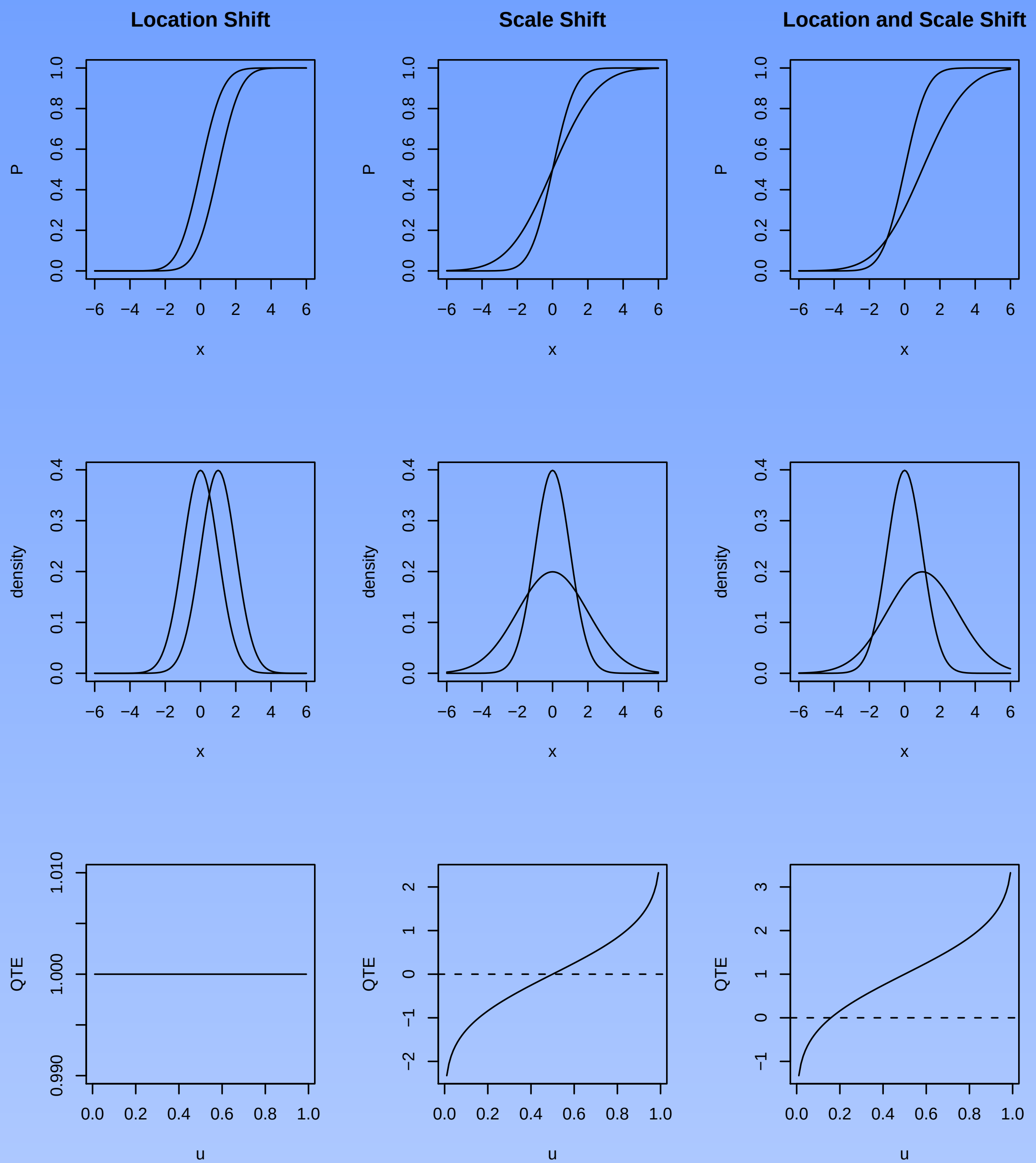
$$\Delta(x) = G^{-1}(F(x)) - x.$$

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

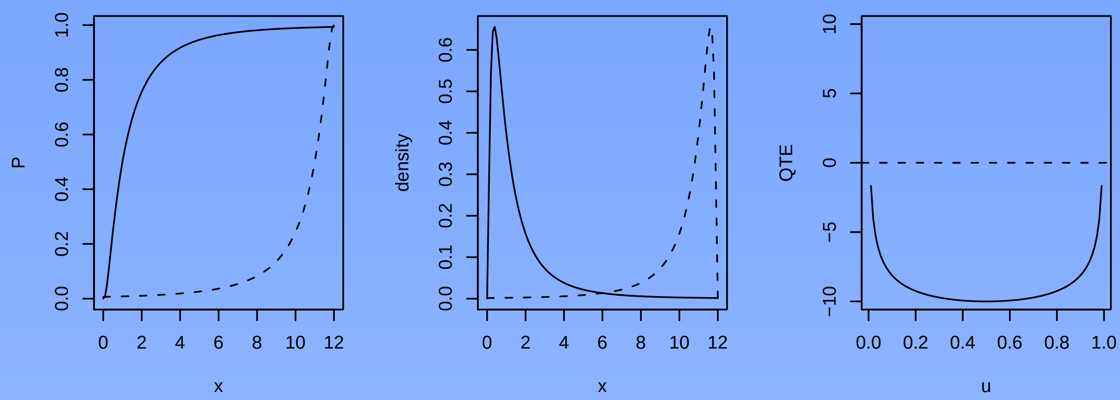
$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$

Lehmann-Doksum QTE


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


Location Shift
Scale Shift
Location and Scale Shift
P
density
QTE
x
u

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,

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

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.

Equivariance of Regression Quantiles

- Scale Equivariance: For any $a > 0$, $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$ and $\hat{\beta}(\tau; -ay, X) = a\hat{\beta}(1 - \tau; y, X)$

Equivariance of Regression Quantiles

- Scale Equivariance: For any $a > 0$, $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$ and $\hat{\beta}(\tau; -ay, X) = a\hat{\beta}(1-\tau; y, X)$

- Regression Shift: For any $\gamma \in \mathbb{R}^p$ $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma$

Equivariance of Regression Quantiles

- Scale Equivariance: For any $a > 0$, $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$ and $\hat{\beta}(\tau; -ay, X) = a\hat{\beta}(1 - \tau; y, X)$

- Regression Shift: For any $\gamma \in \mathbb{R}^p$ $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma$

- Reparameterization of Design: For any $|A| \neq 0$, $\hat{\beta}(\tau; y, AX) = A^{-1}\hat{\beta}(\tau; yX)$

Equivariance to Monotone Transformations

For any monotone function h, conditional quantile functions $Q_Y(\tau|x)$ are equivariant in the sense that

$$Q_{h(Y)|X}(\tau|x) = h(Q_{Y|X}(\tau|x))$$

In contrast to conditional mean functions for which

$$E(h(Y)|X) \neq h(EY|X)$$

Examples:

$h(y) = \min\{0, y\}$, Powell(1985)

$h(y) = \mathrm{sgn}\{y\}$ Rosenblatt(1957) Manski(1975)

Robustness

- Bounded Influence Function in y for fixed x_i, decent breakdown behavior for fixed design.

- Only the signs of the residuals $\hat{u} = y - X\hat{\beta}(\tau, y, X)$. matter

$$\hat{\beta}(\tau; y, X) = \hat{\beta}(\tau, y + D\hat{u}, X)$$

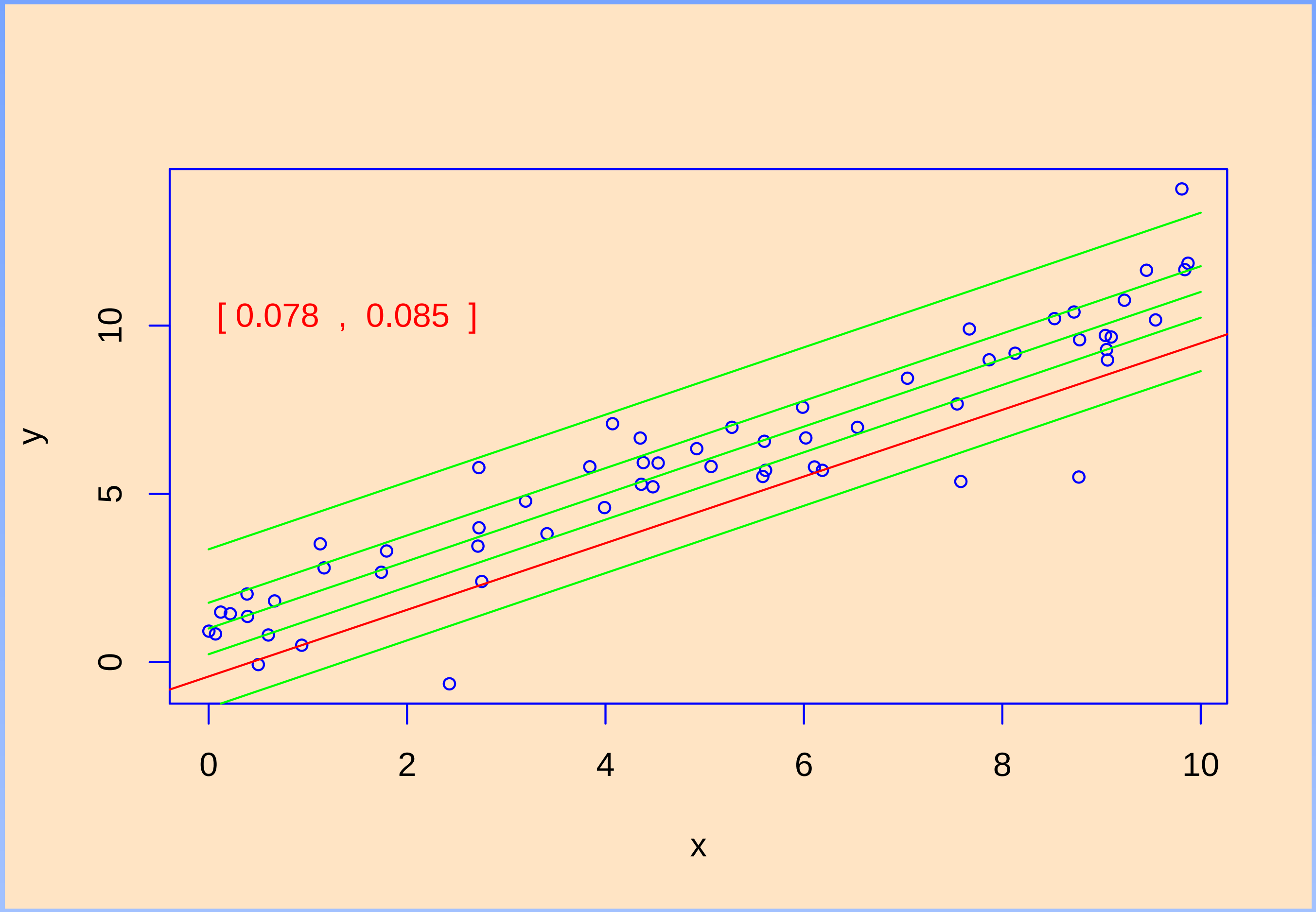
  for any diagonal matrix D with nonnegative elements.

- Robustness with respect to influential x observations is more challenging, but there are several very interesting proposals.

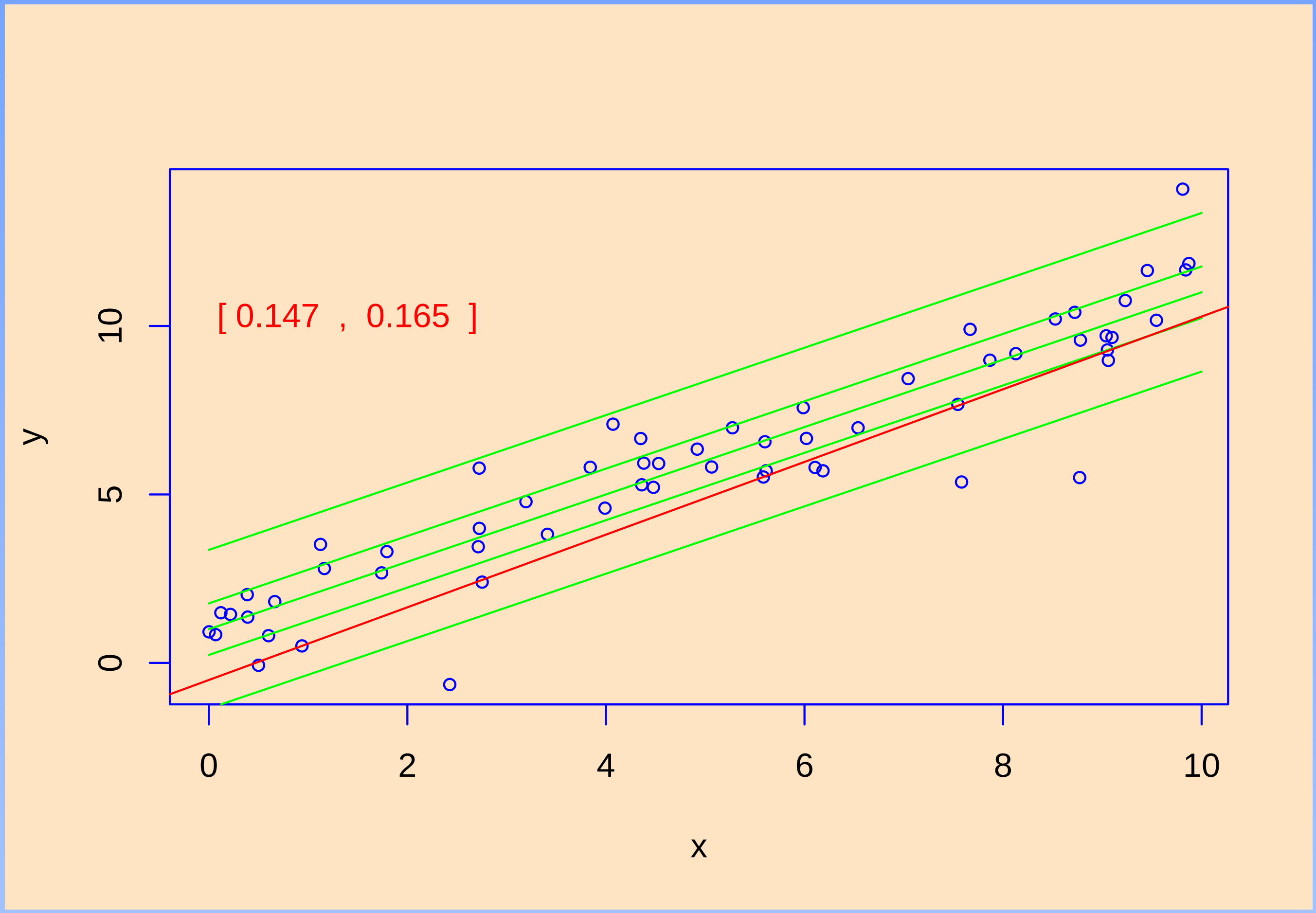
Quantile Regression: The Movie

- 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

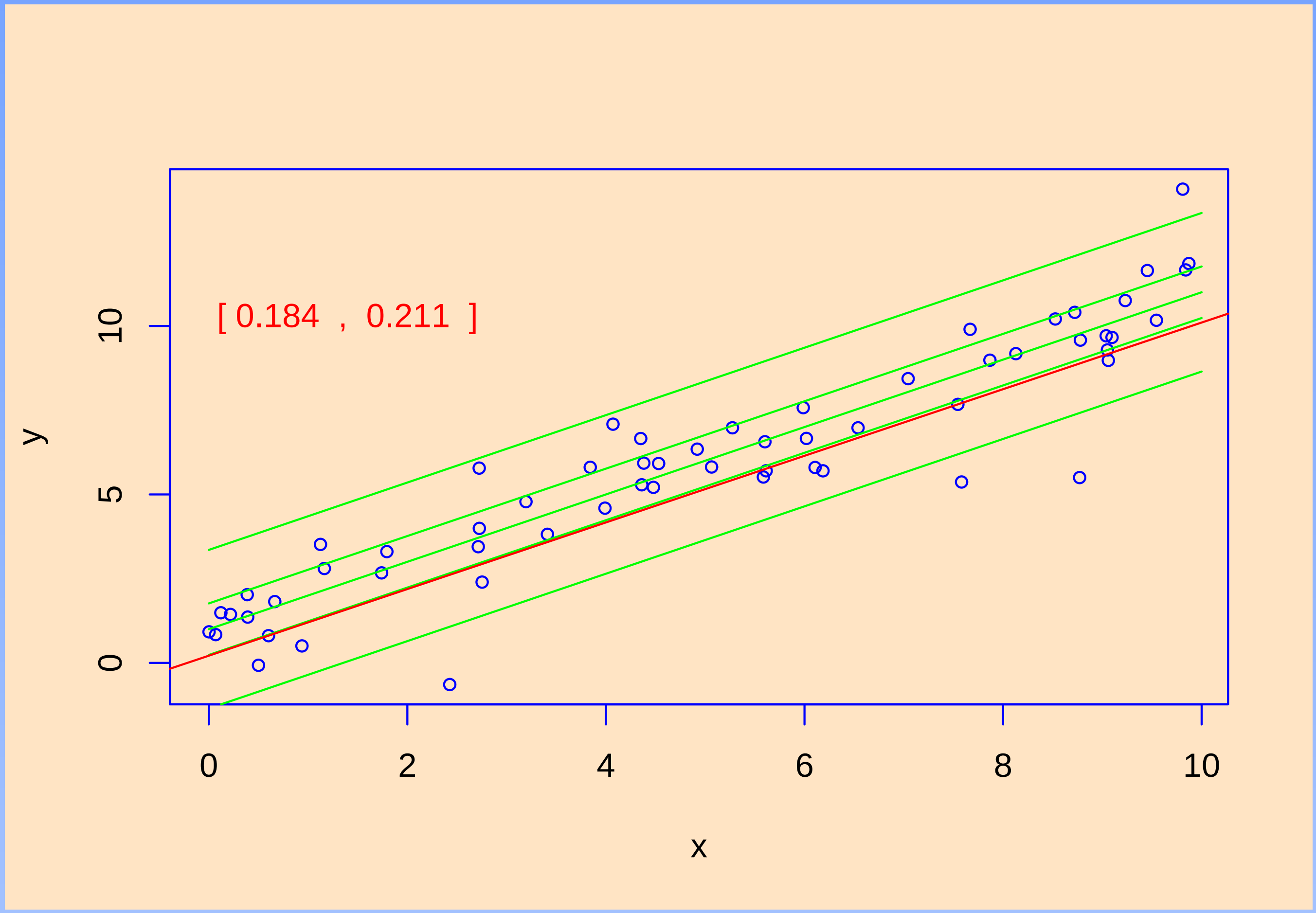
Quantile Regression in the iid Error Model



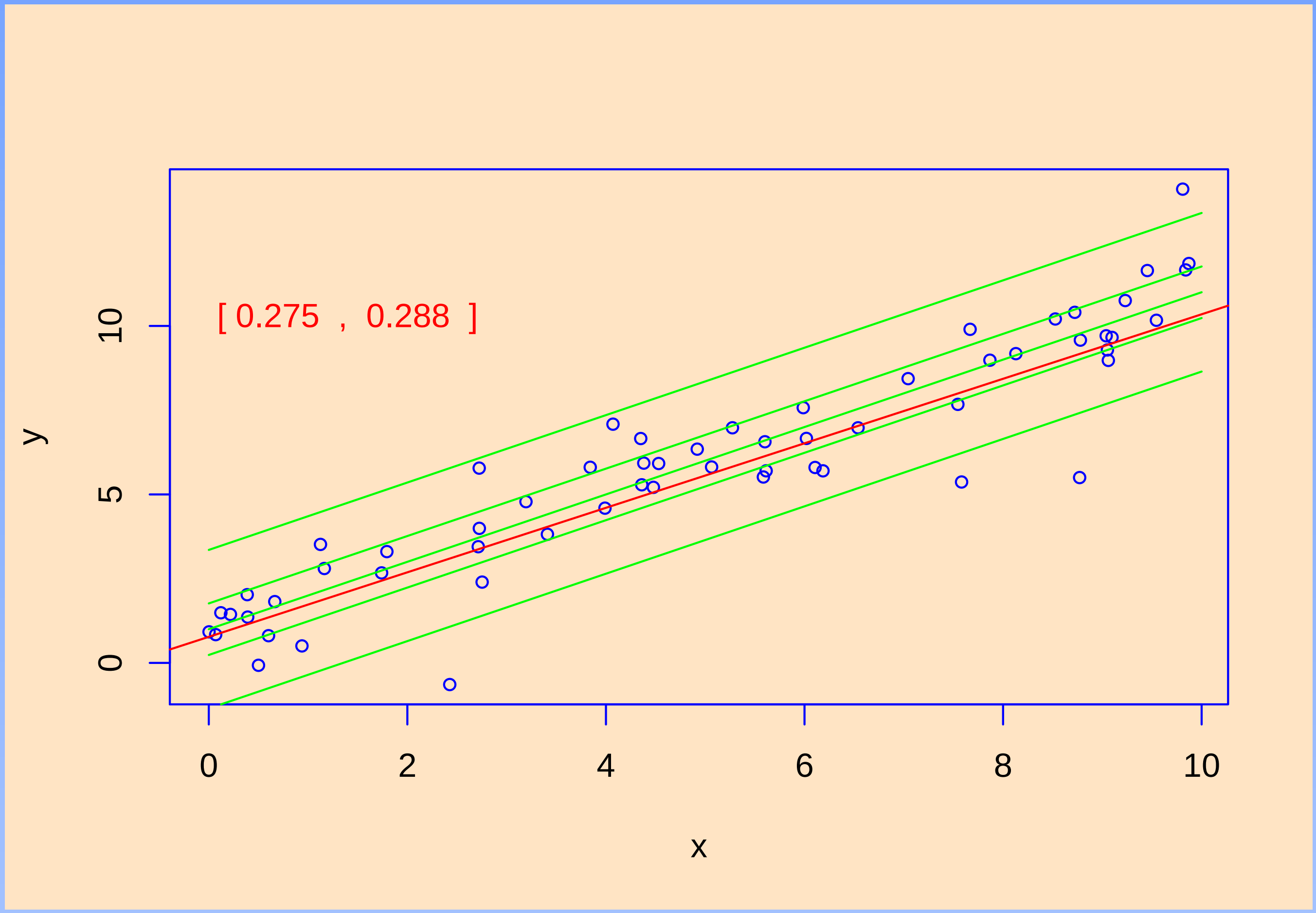
Quantile Regression in the iid Error Model



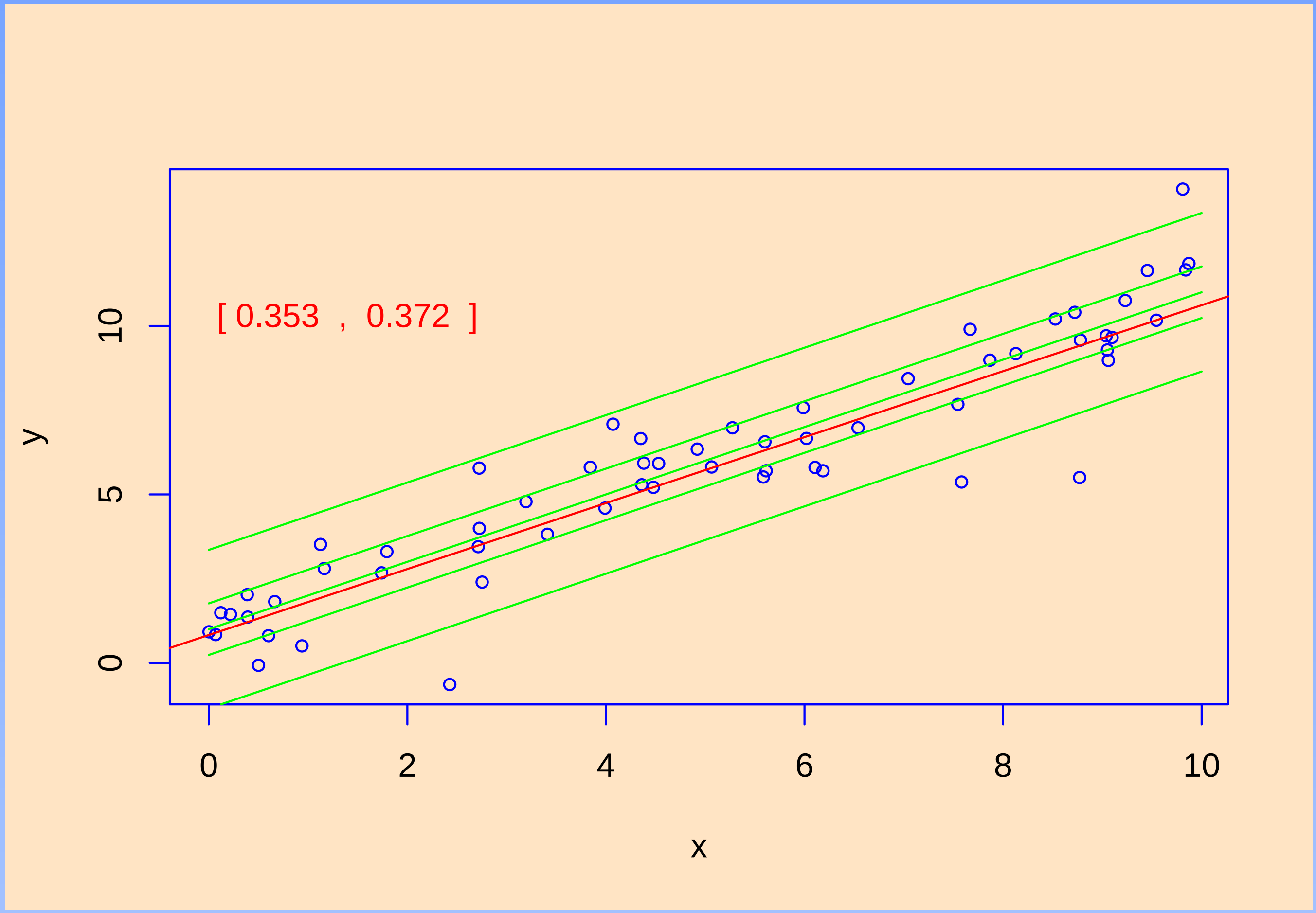
Quantile Regression in the iid Error Model



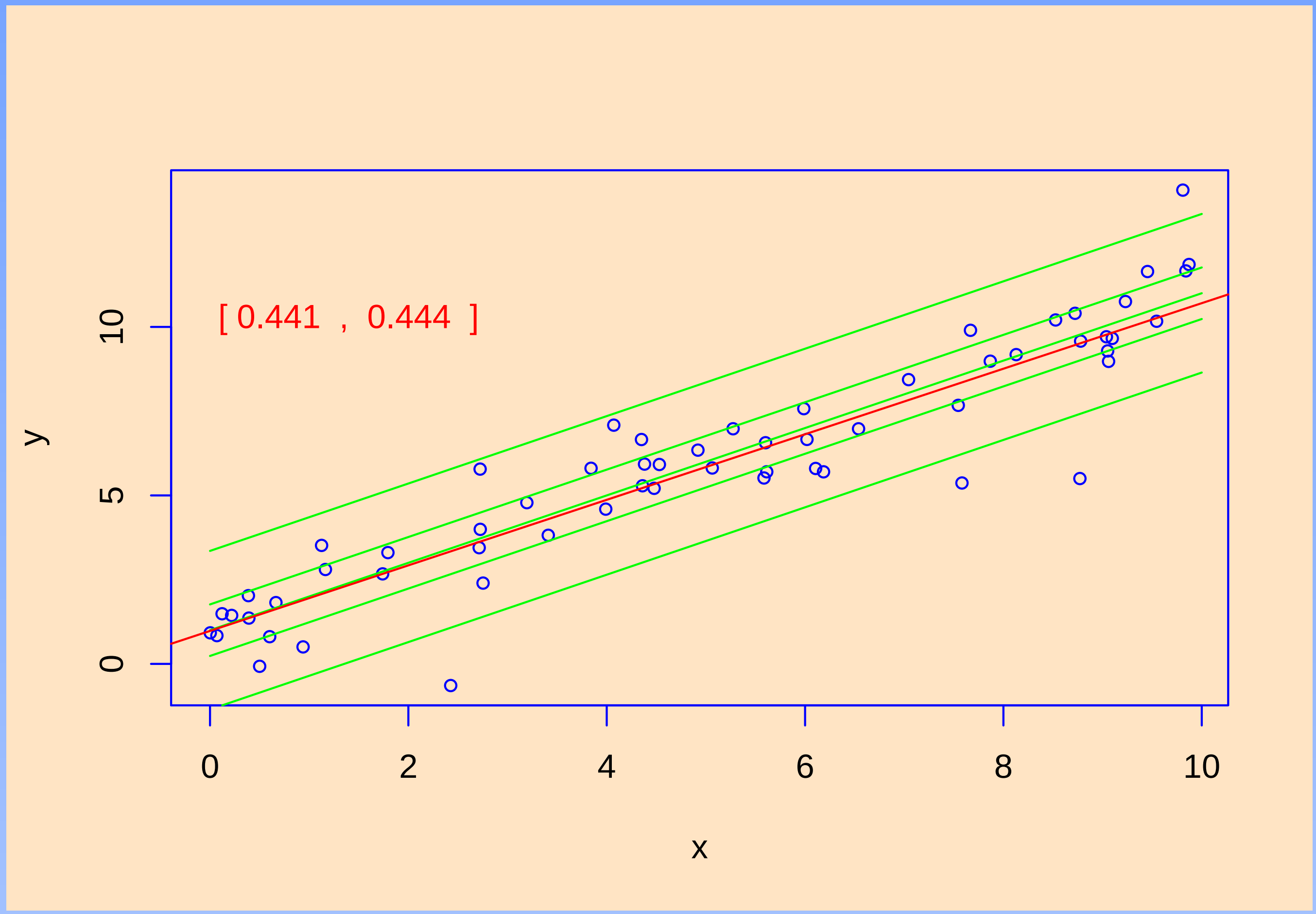
Quantile Regression in the iid Error Model



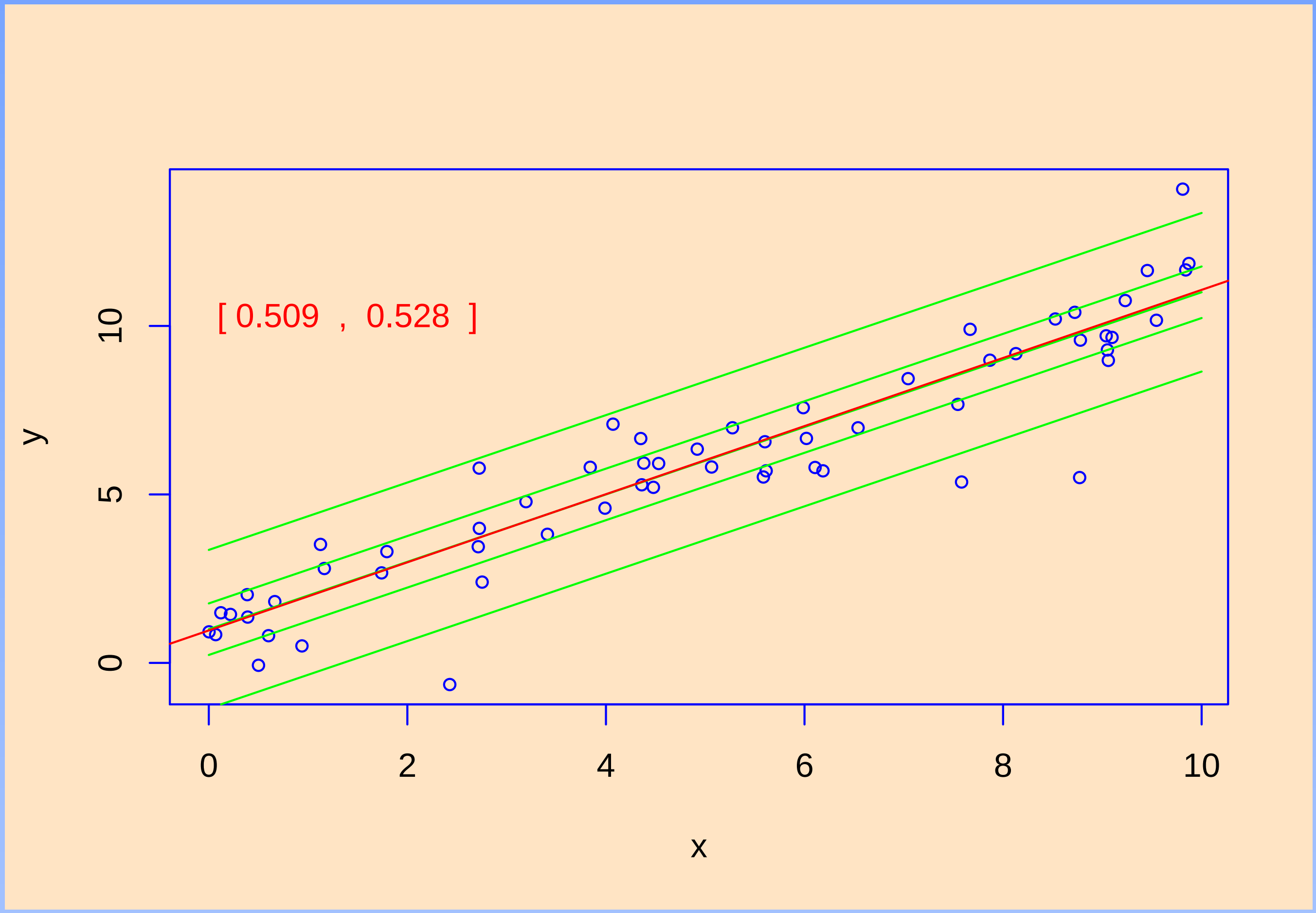
Quantile Regression in the iid Error Model



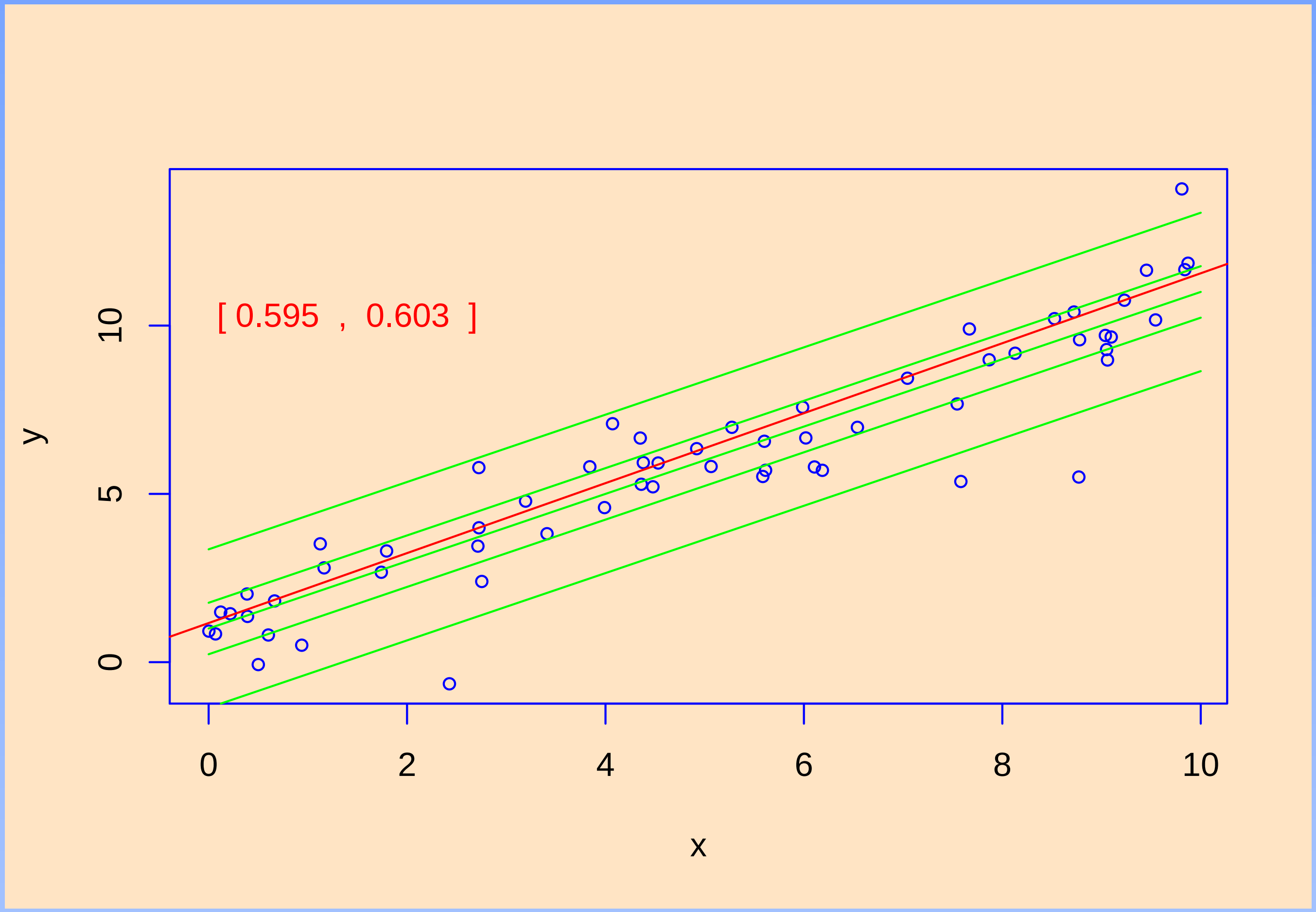
Quantile Regression in the iid Error Model



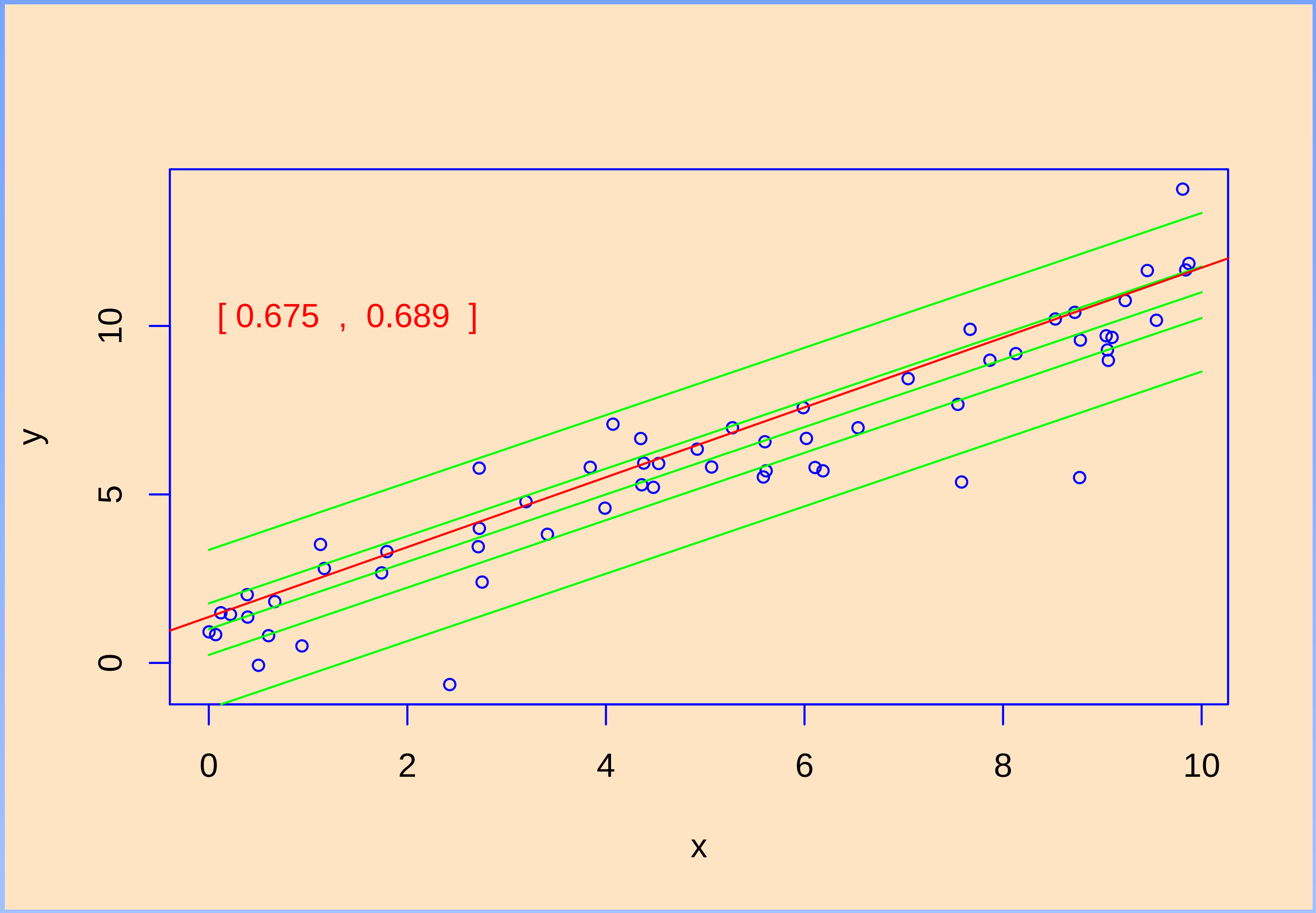
Quantile Regression in the iid Error Model



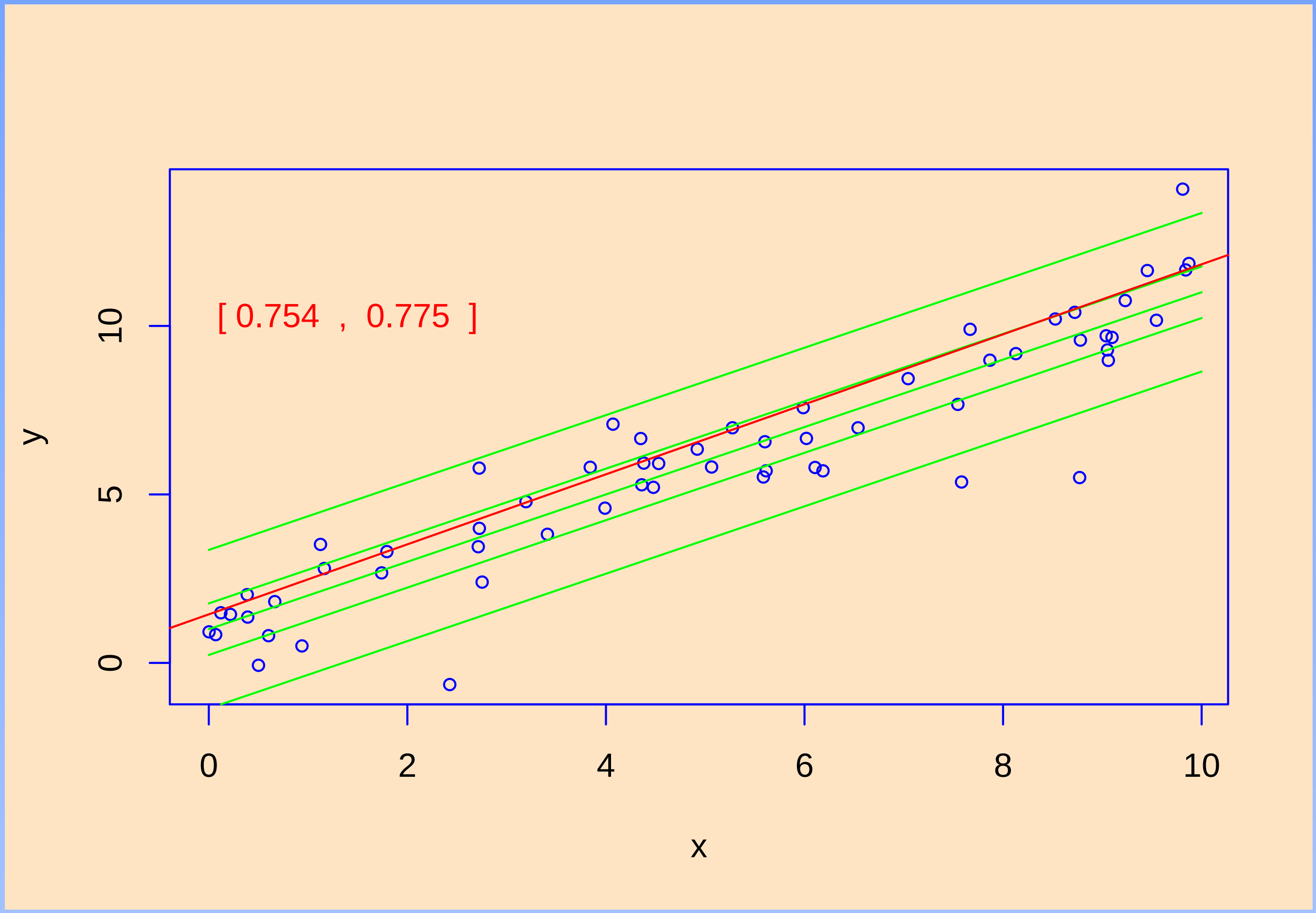
Quantile Regression in the iid Error Model



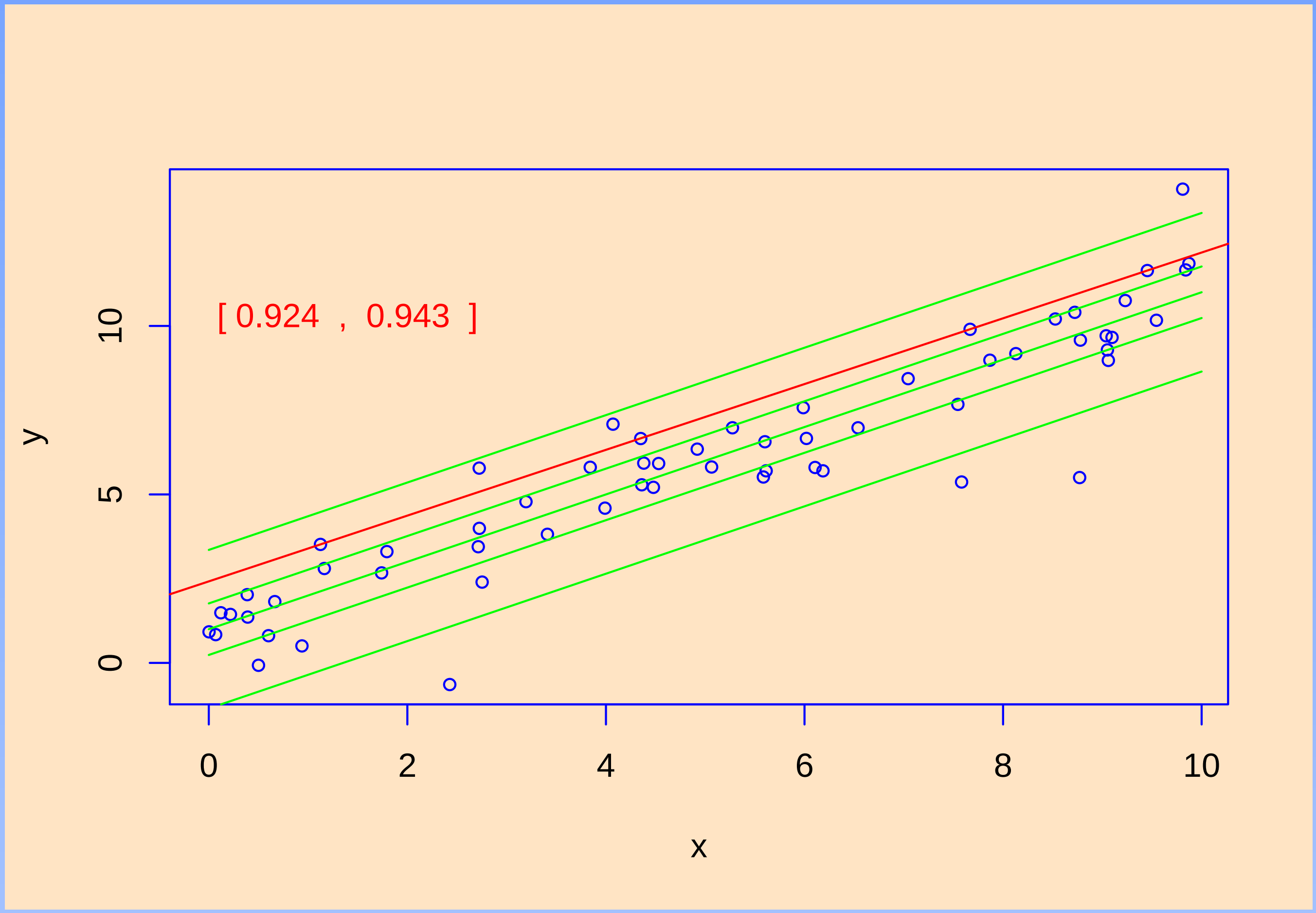
Quantile Regression in the iid Error Model



Quantile Regression in the iid Error Model



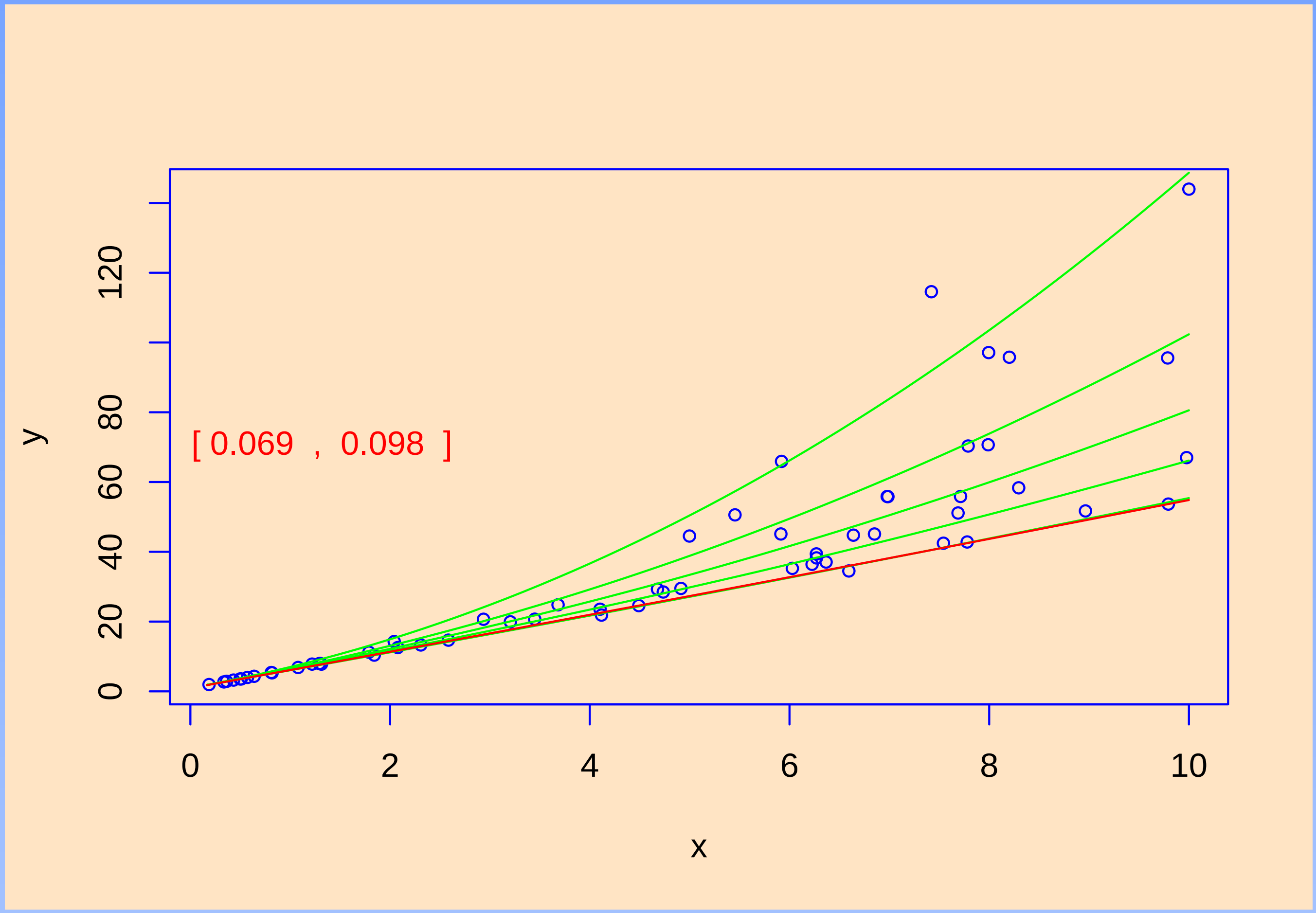
Quantile Regression in the iid Error Model



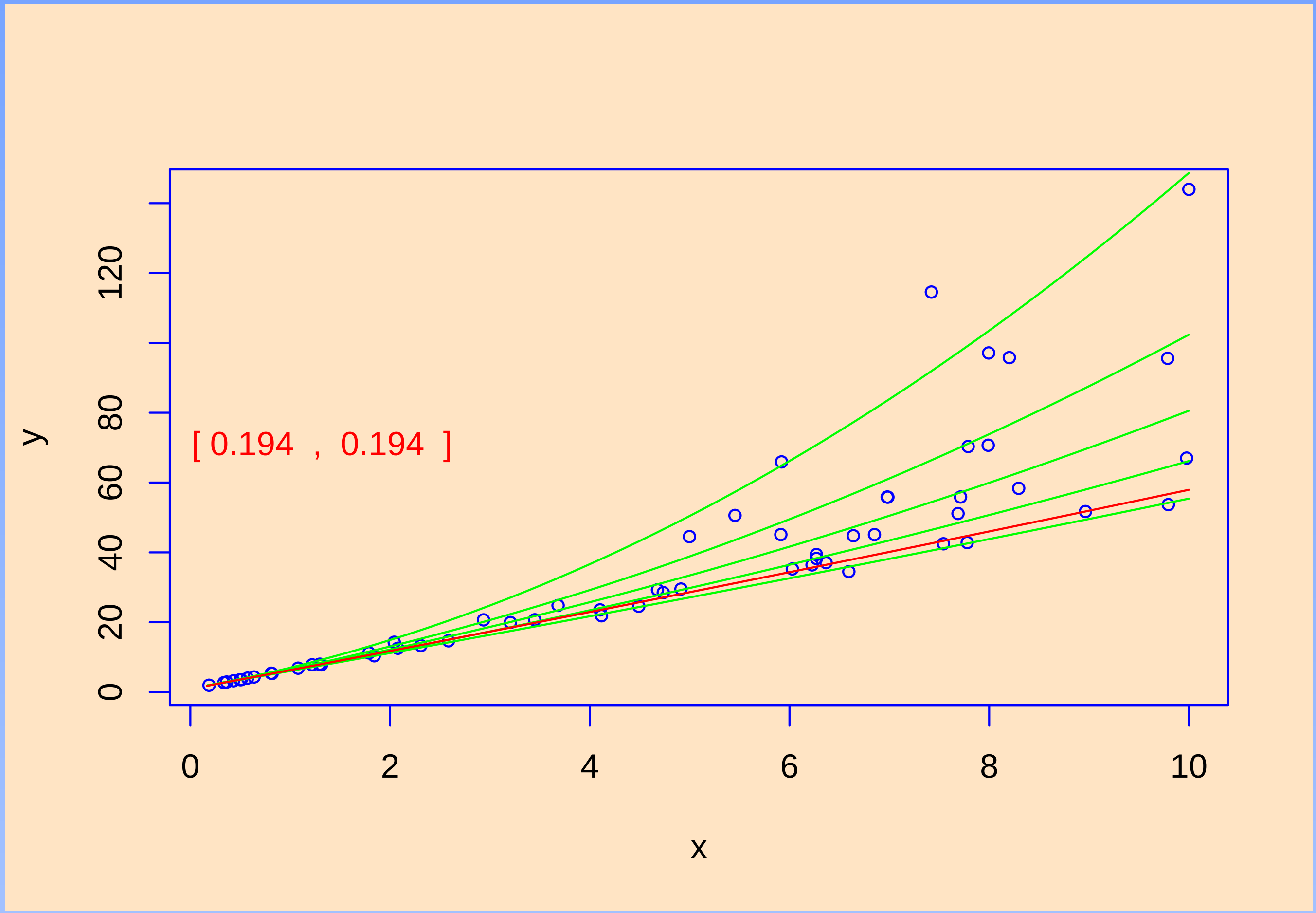
Virtual Quantile Regression II

- 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

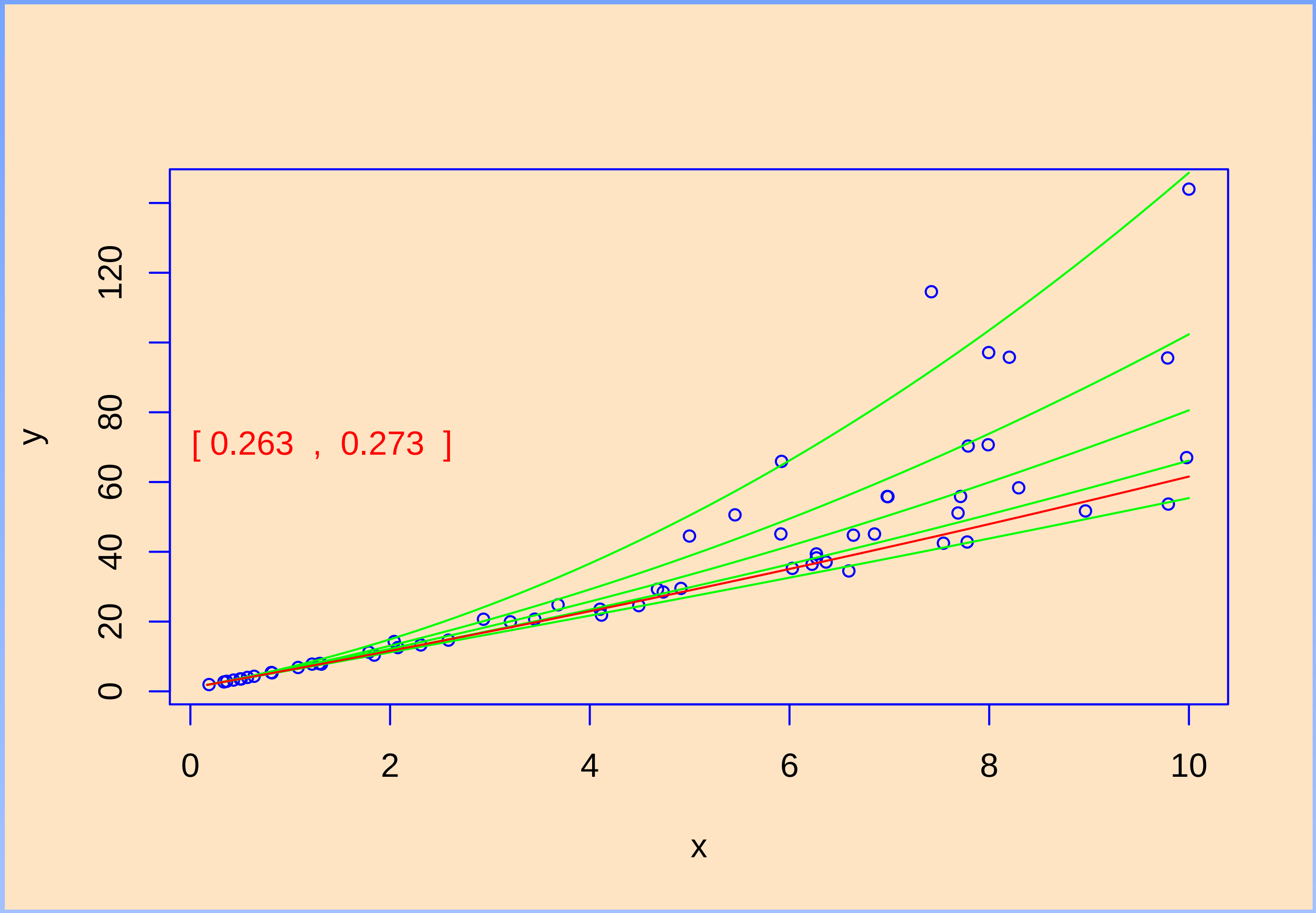
Quantile Regression in the Heteroscedastic Error Model



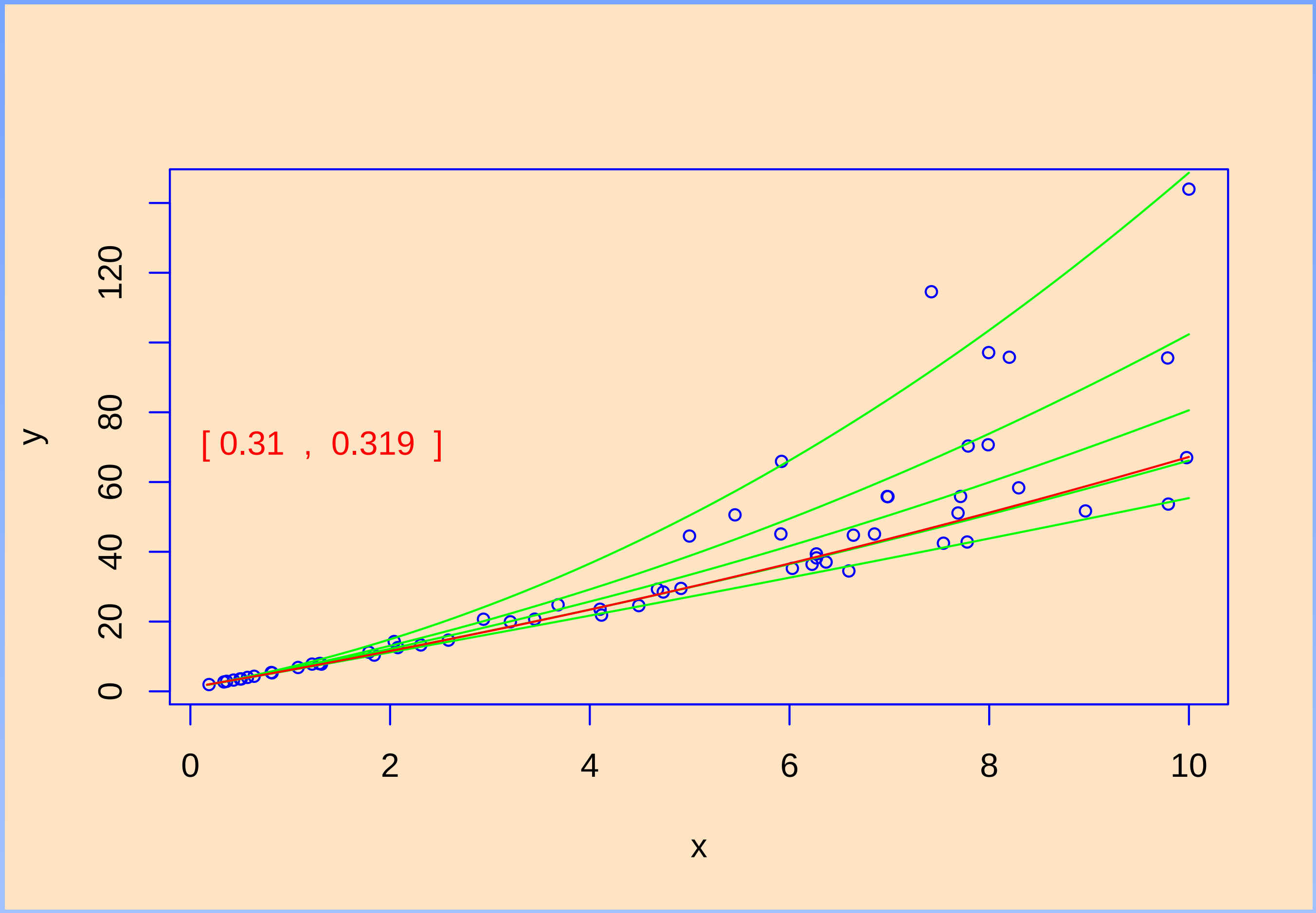
Quantile Regression in the Heteroscedastic Error Model



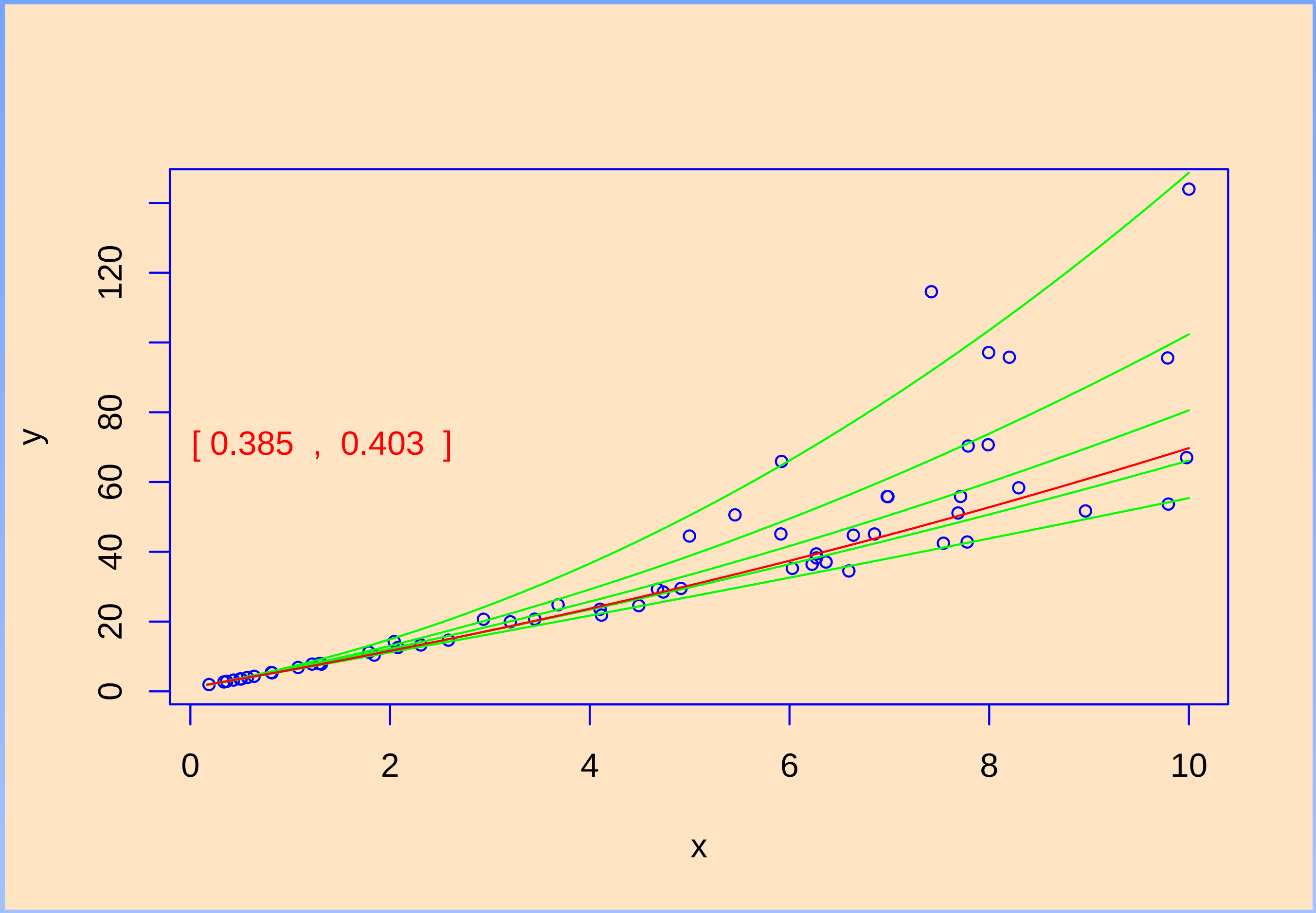
Quantile Regression in the Heteroscedastic Error Model



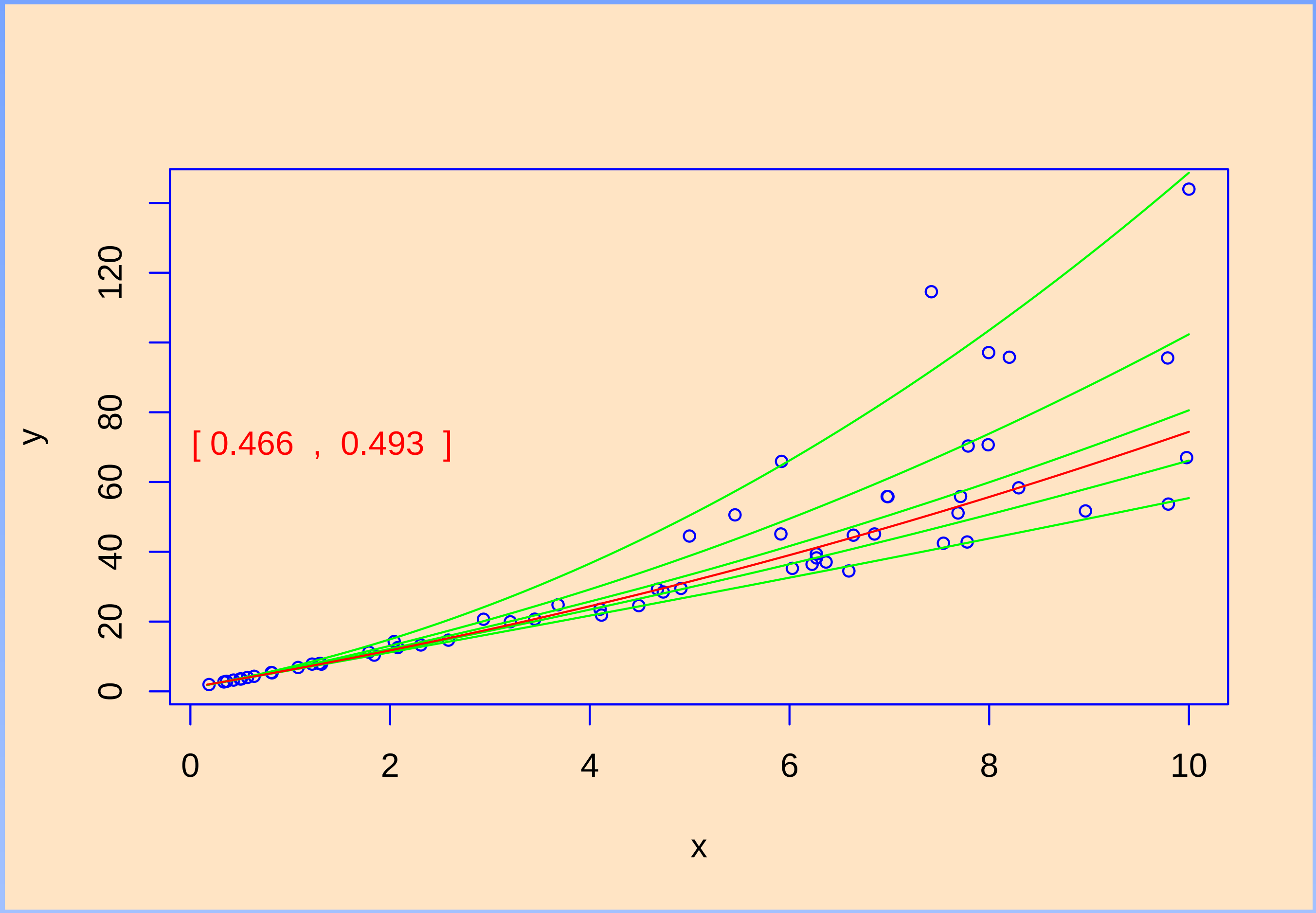
Quantile Regression in the Heteroscedastic Error Model



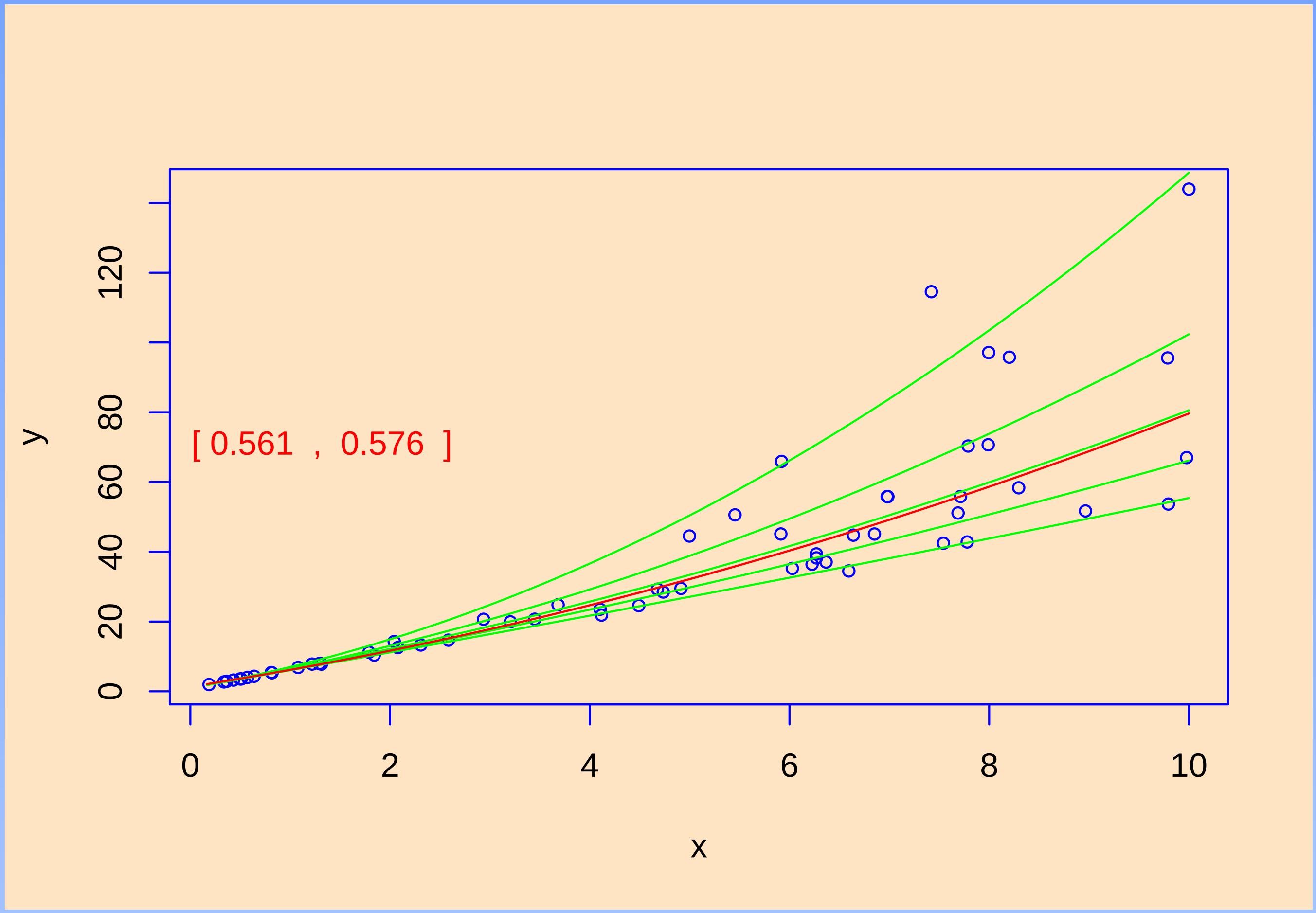
Quantile Regression in the Heteroscedastic Error Model



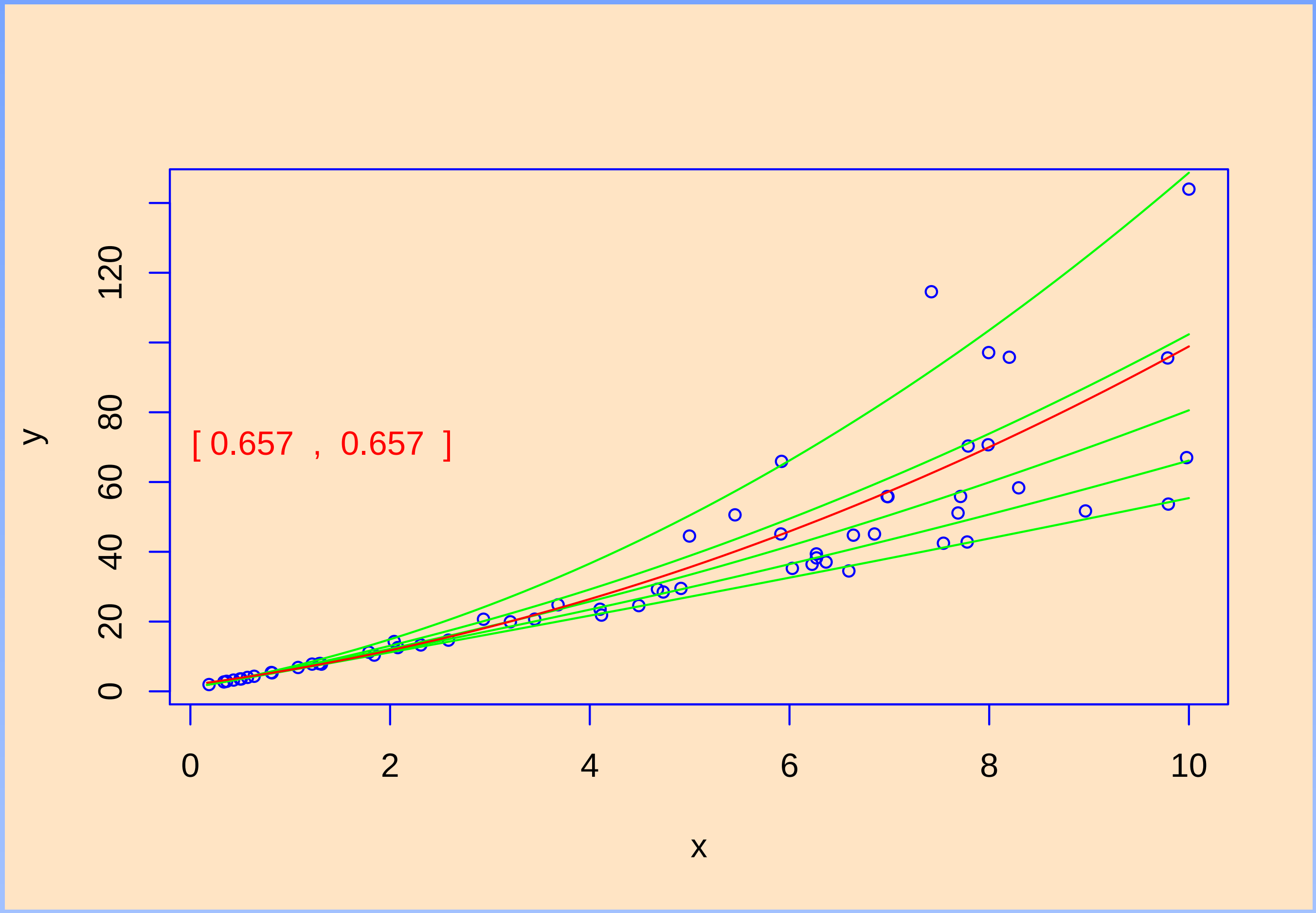
Quantile Regression in the Heteroscedastic Error Model



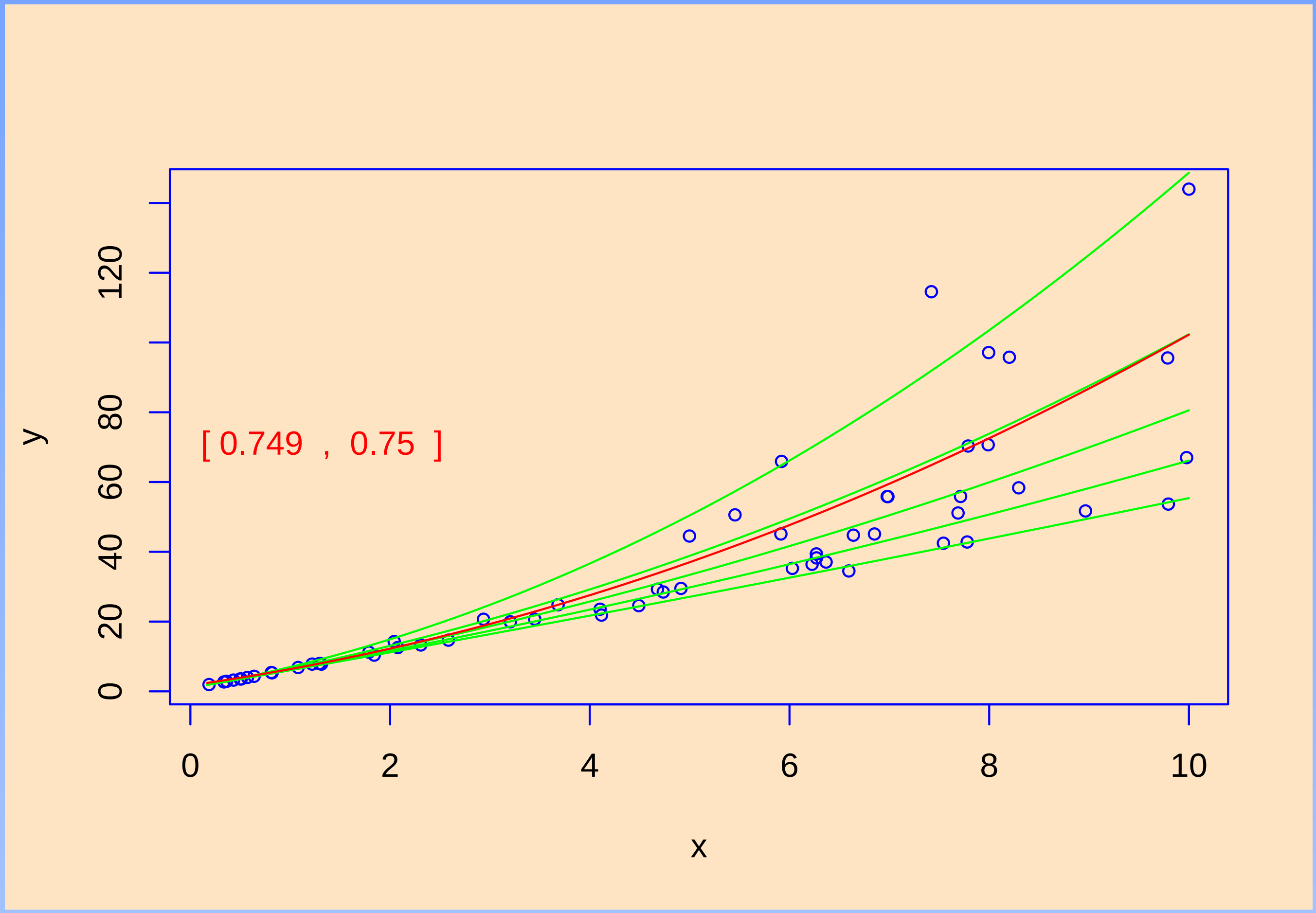
Quantile Regression in the Heteroscedastic Error Model



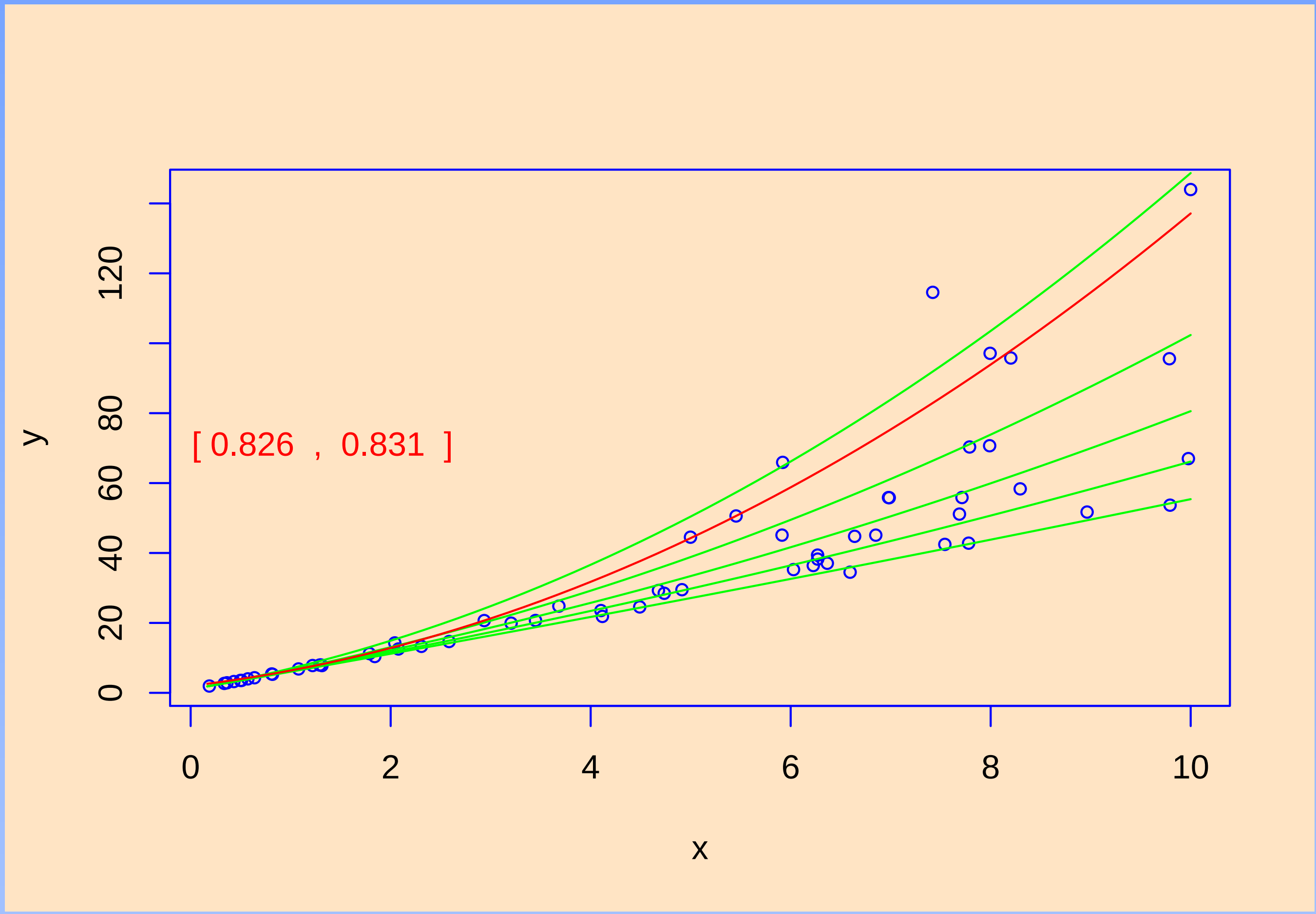
Quantile Regression in the Heteroscedastic Error Model



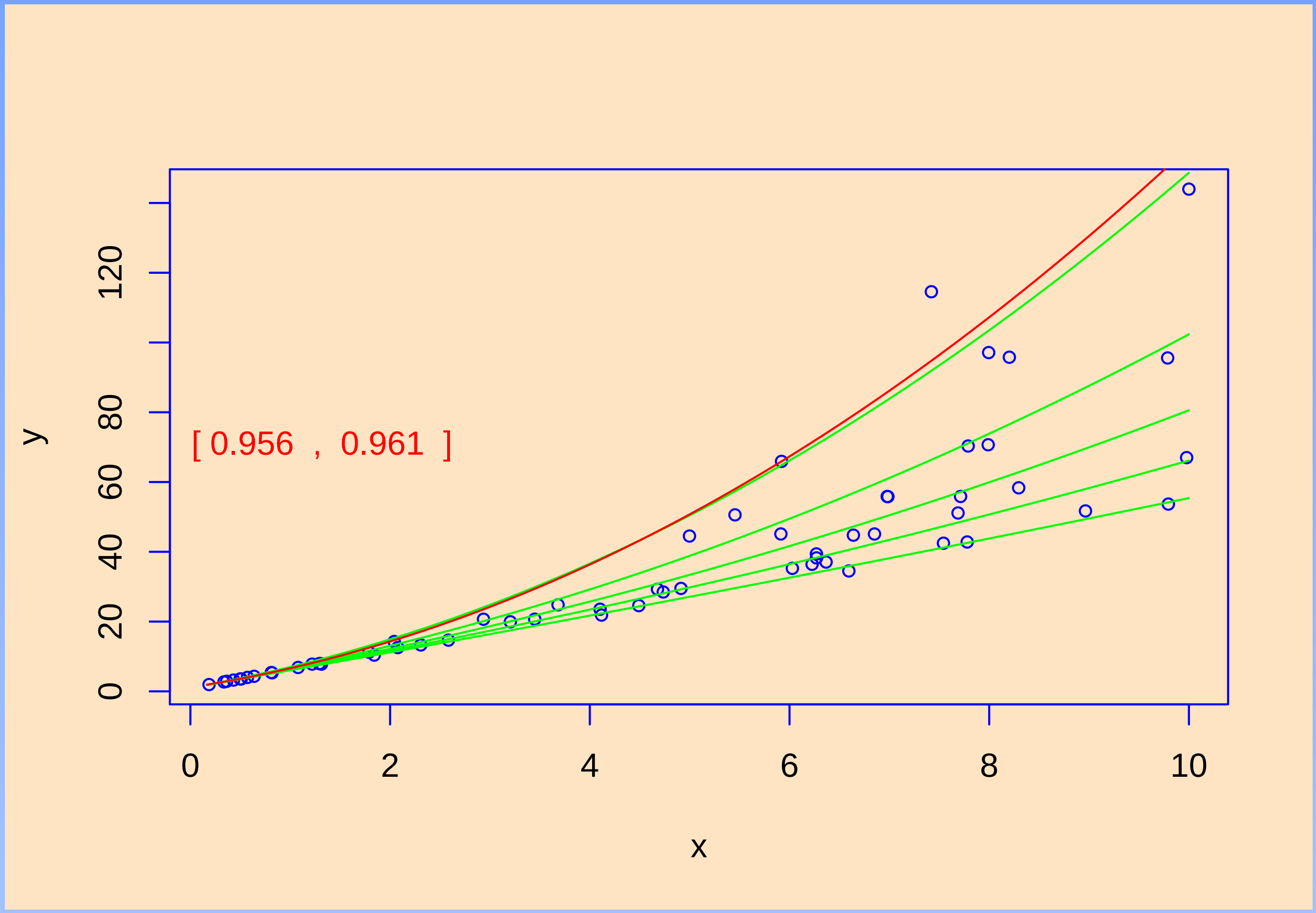
Quantile Regression in the Heteroscedastic Error Model



Quantile Regression in the Heteroscedastic Error Model



Quantile Regression in the Heteroscedastic Error Model



Three Applications

- Engel's Law: A Classical Economic Example

- Infant Birthweight: A Public Health Example

- Melbourne Daily Temperature: A Time Series Example

Engel's Food Expenditure Data

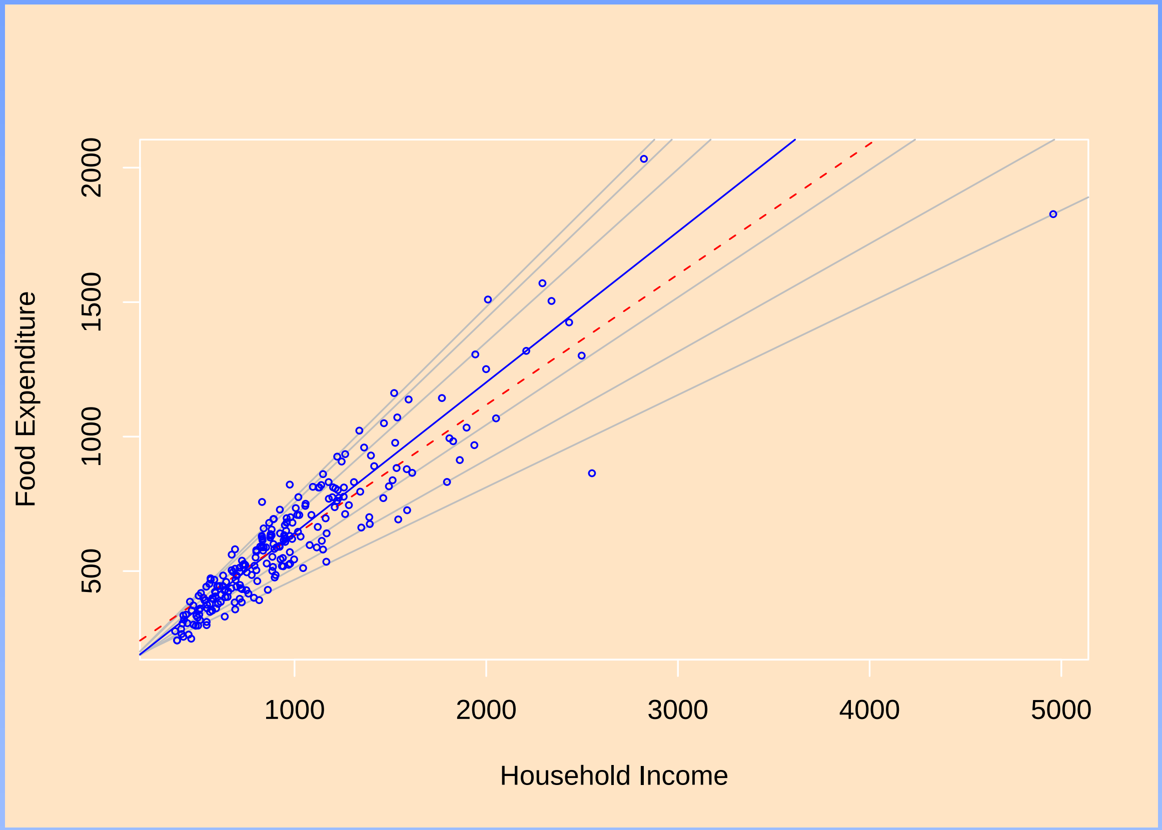


Figure 1: Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ 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

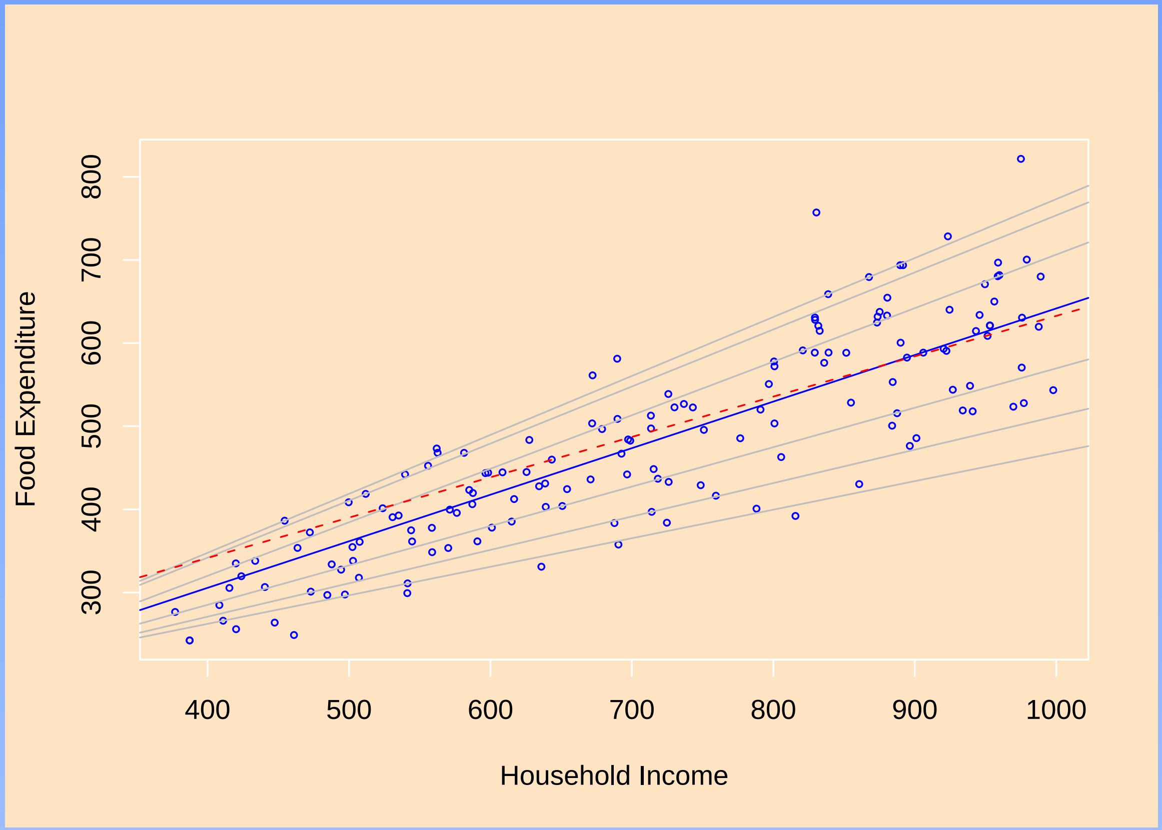
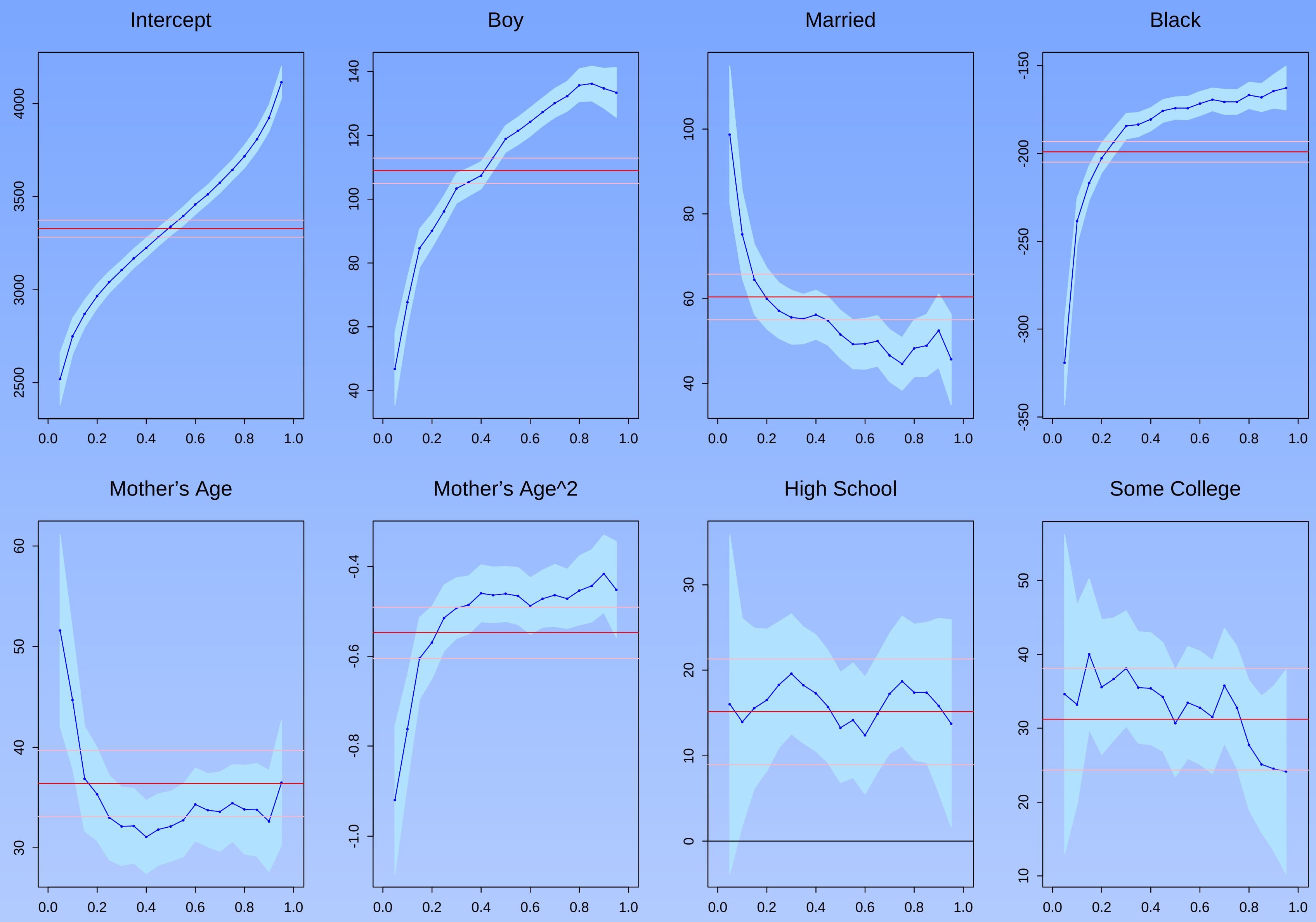


Figure 2: Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau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

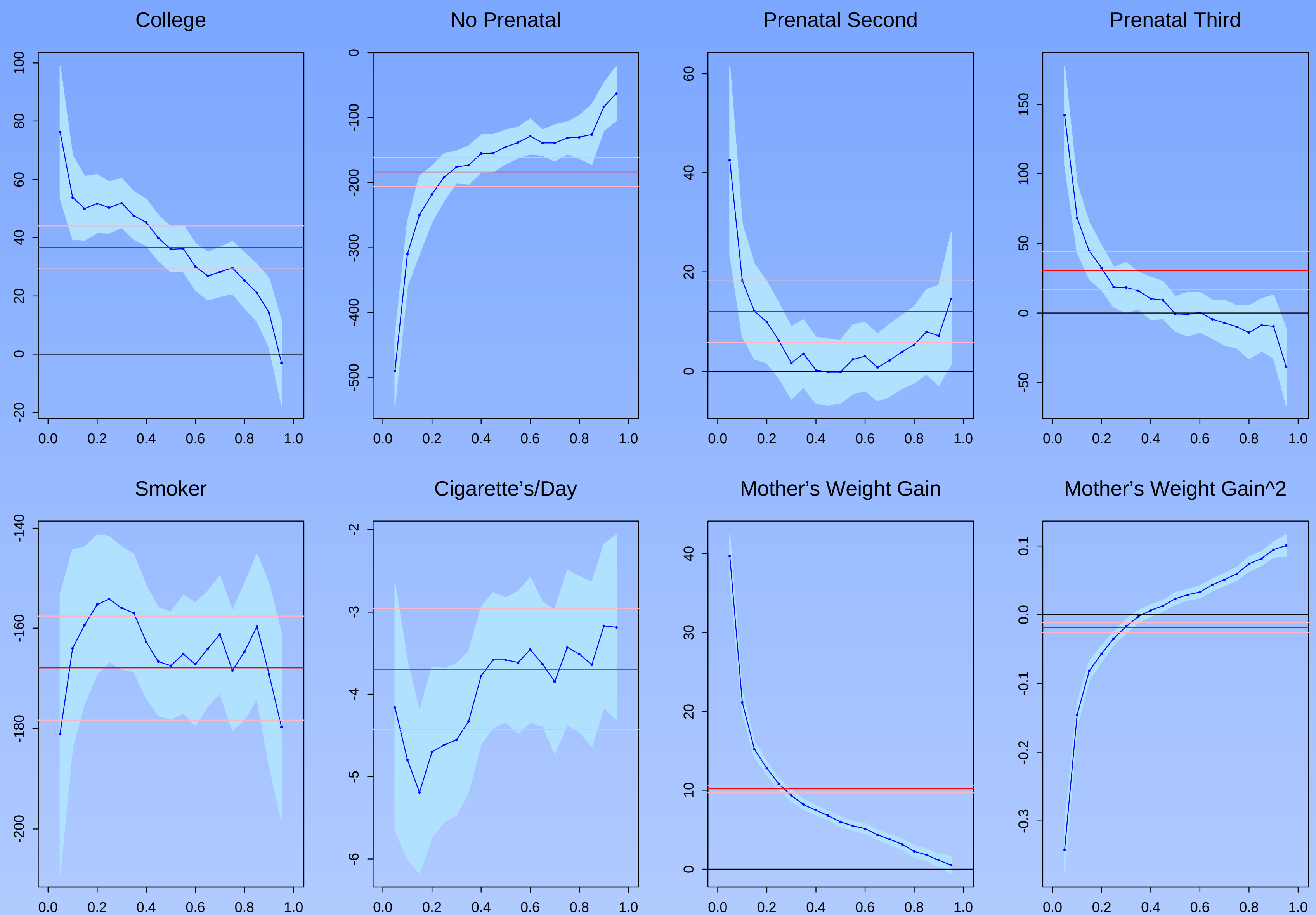
A Model of Infant Birthweight

- 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

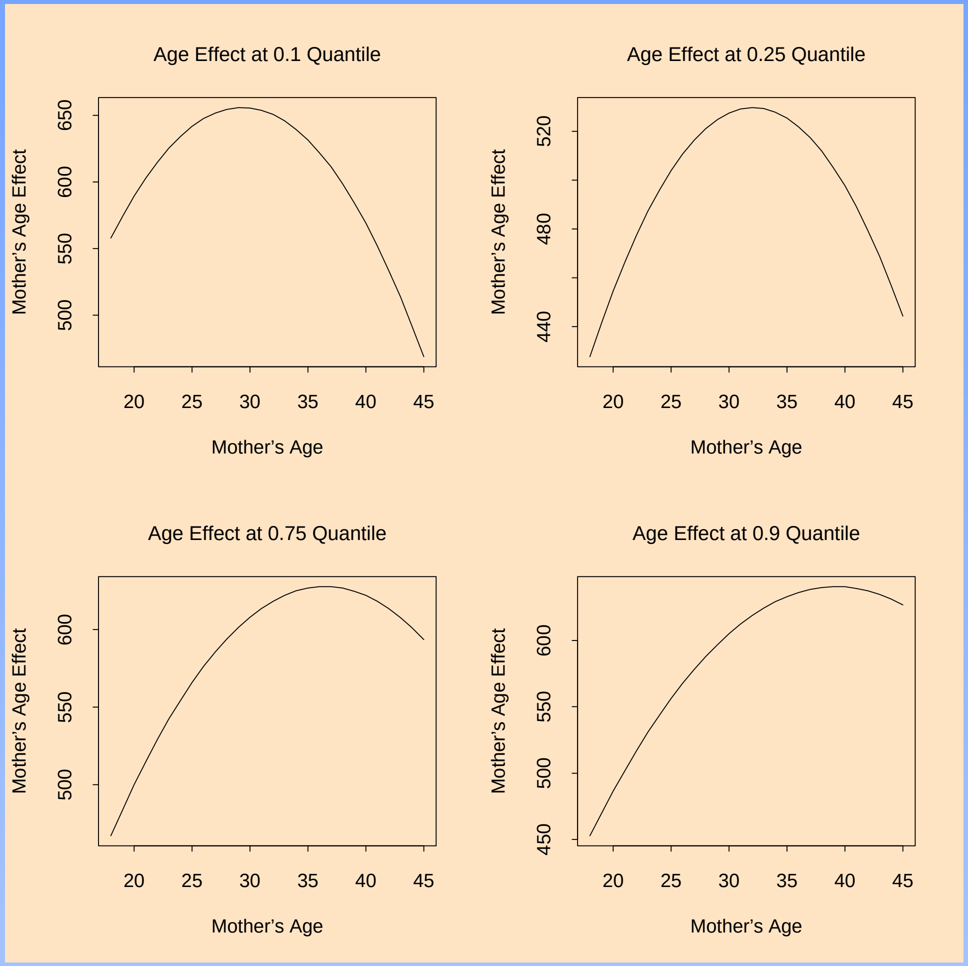
Quantile Regression Birthweight Model I



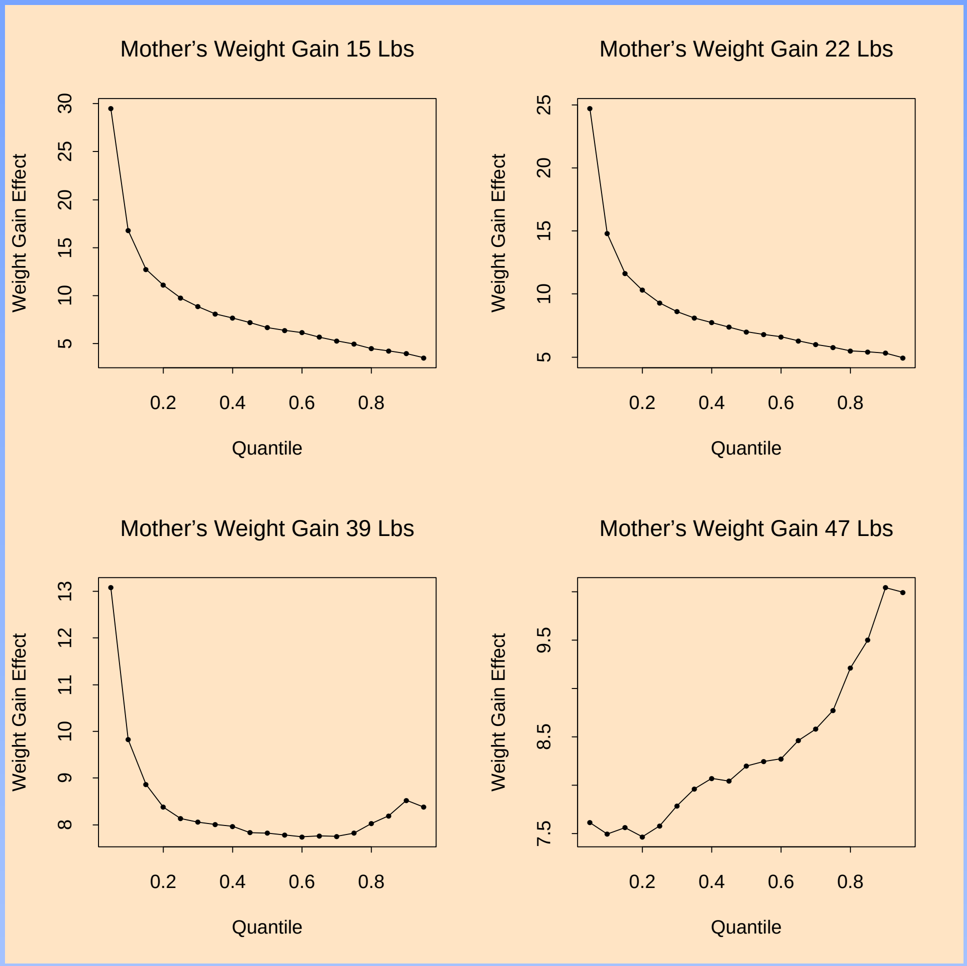
Quantile Regression Birthweight Model II



Marginal Effect of Mother's Age



Marginal Effect of Mother's Weight Gain



AR(1) Model of Melbourne Daily Temperature

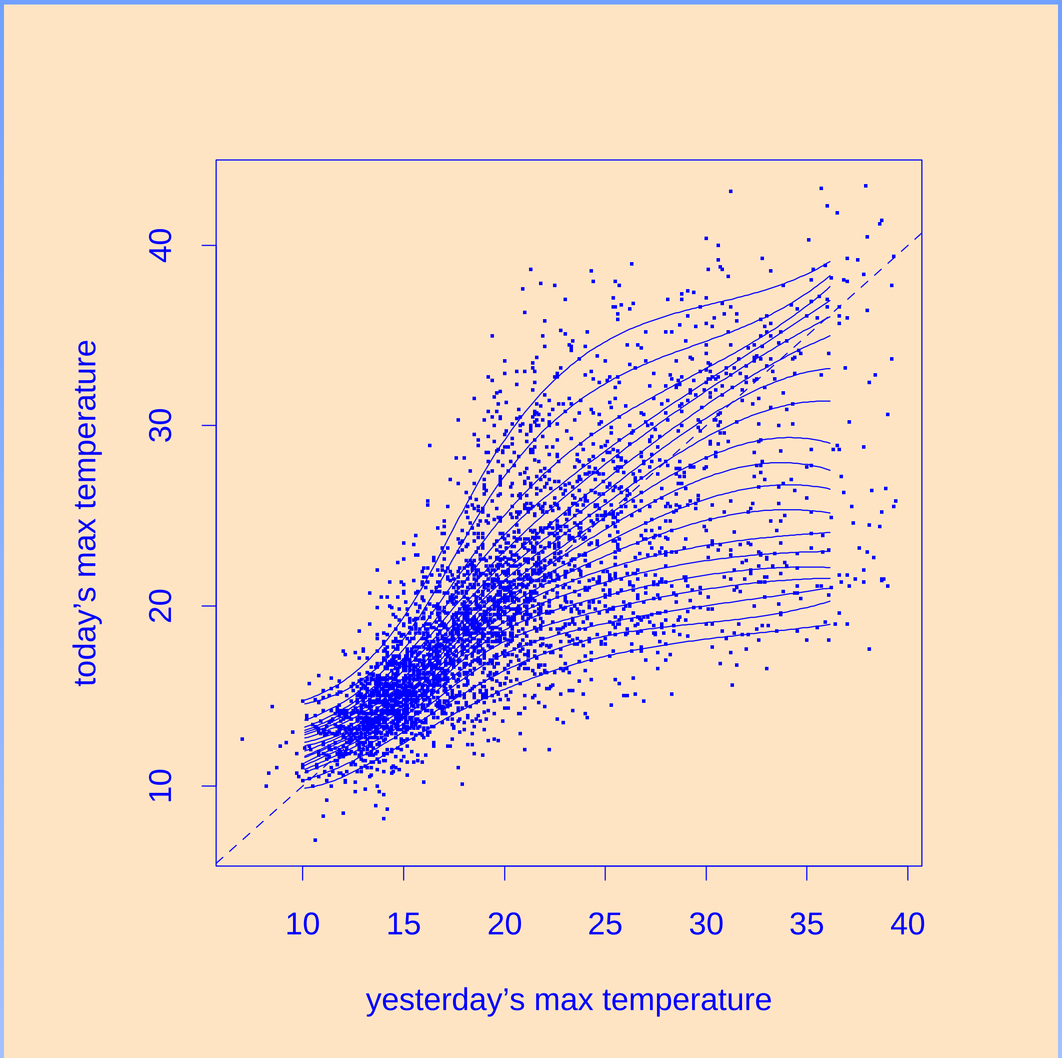
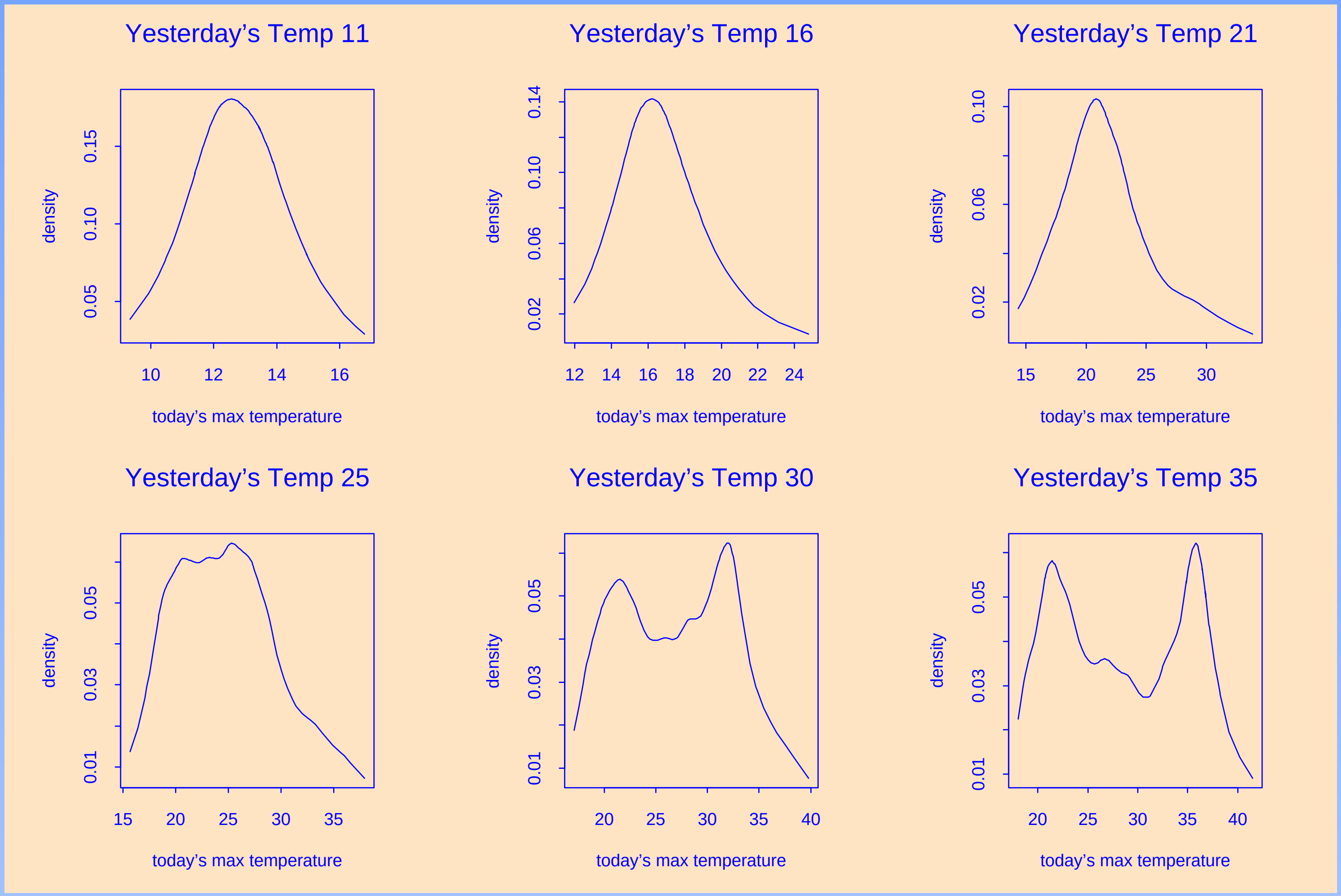


Figure 3: The plot illustrates 10 years of daily maximum temperature data for Melbourne, Australia as an AR(1) scatterplot. Superimposed are estimated conditional quantile functions for $\tau \in \{.05, .10, ..., .95\}$.

Conditional Densities of Melbourne Daily Temperature



Conclusions

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

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.

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.

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.