Quantile Regression: A Gentle Introduction

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

Les Diablerets 3-6 February 2013

Overview of the Lectures

- The Basics: What, Why and How?
- Inference and Quantile Treatment Effects
- Nonparametric Quantile Regression
- Endogoneity and IV Methods
- Censored QR and Survival Analysis
- Quantile Autoregression
- QR for Longitudinal Data
- Risk Assessment and Choquet Portfolios
- Computional Aspects

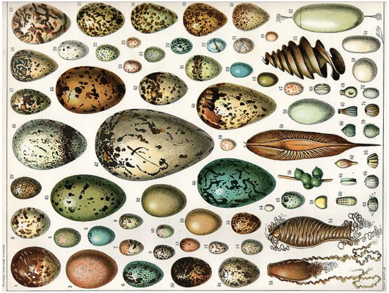
Course outline, lecture slides, an R FAQ, and even some proposed exercises can all be found at:

`http://www.econ.uiuc.edu/~roger/courses/Diab`.

The Basics: What, Why and How?

1. Univariate Quantiles
2. Scatterplot Smoothing
3. Equivariance Properties
4. Quantile Treatment Effects
5. Three Empirical Examples

Archimedes' "Eureka!" and the Middle Sized Egg



Volume of the eggs can be measure by the amount of water they displace and the median (middle-sized) egg found by sorting these measurements.

Note that even if we measure the logarithm of the volumes, the middle sized egg is the same. Not true for the mean egg!

The Stem and Leaf Plot: Tukey's EDA Gadget Number 1

Given a "batch" of numbers, $\{X_1, X_2, ..., X_n\}$ one can make a quick and dirty histogram in R this way:

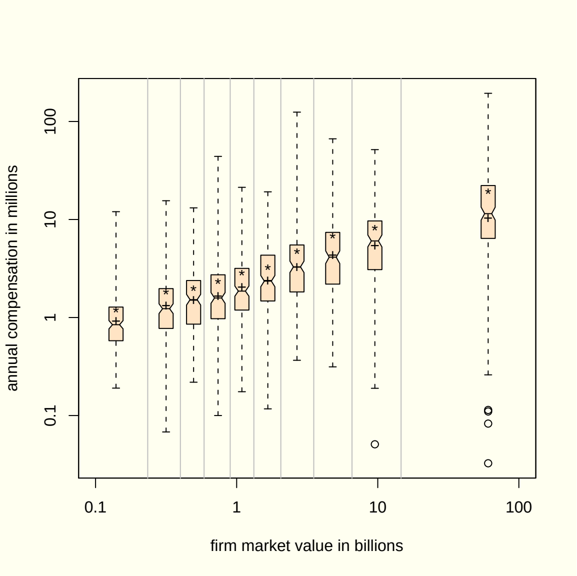
```
> x <- rchisq(100,5) # 100 Chi-squared(5)
> quantile(x) # Tukey's Five Number Summary
        0%        25%        50%        75%       100%
 0.9042396  2.7662230  4.2948642  6.2867588 16.5818573

> stem(x)

  The decimal point is at the |

   0 | 92356668
   2 | 001111244445667778889990111222455666
   4 | 01223334666678901125567889
   6 | 023344667802888
   8 | 556691
  10 | 7
  12 | 159
  14 | 06
  16 | 6
```

Boxplot of CEO Pay: Tukey's EDA Gadget Number 2



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.

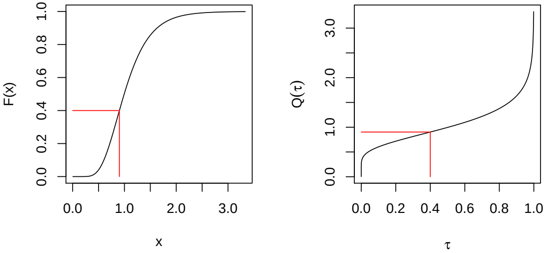
Mosteller and Tukey (1977)

Univariate Quantiles

Given a real-valued random variable, X, with distribution function F, we will define the τth quantile of X as

$$Q_X(\tau) = F_X^{-1}(\tau) = \inf\{x \mid F(x) \geqslant \tau\}.$$

This definition follows the usual convention that F is CADLAG, and Q is CAGLAD as illustrated in the following pair of pictures.

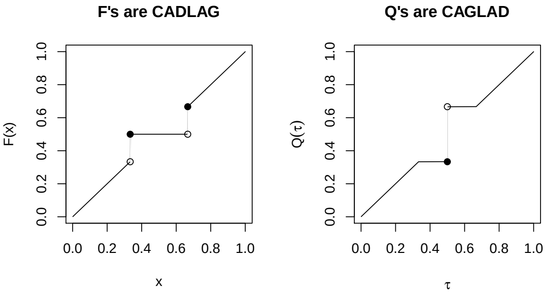


Univariate Quantiles

Given a real-valued random variable, X, with distribution function F, we will define the τth quantile of X as

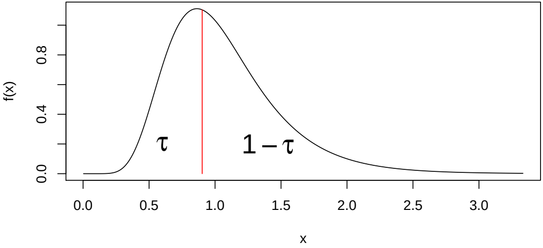
$$Q_X(\tau) = F_X^{-1}(\tau) = \inf\{x \mid F(x) \geqslant \tau\}.$$

This definition follows the usual convention that F is CADLAG, and Q is CAGLAD as illustrated in the following pair of pictures.



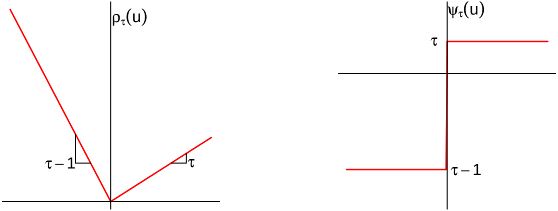
Univariate Quantiles

Viewed from the perspective of densities, the τth quantile splits the area under the density into two parts: one with area τ below the τth quantile and the other with area $1 - \tau$ above it:



Two Bits Worth of Convex Analysis

A convex function ρ and its subgradient ψ:



The subgradient of a convex function $f(u)$ at a point u consists of all the possible "tangents." Sums of convex functions are convex.

Population Quantiles as Optimizers

Quantiles solve a simple optimization problem:

$$\hat{\alpha}(\tau) = \text{argmin}\ \mathbb{E}\ \rho_\tau(Y - \alpha)$$

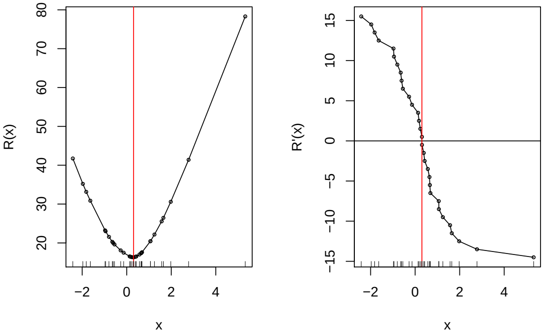
Proof: Let $\psi_\tau(u) = \rho_\tau^{'}(u)$, so differentiating wrt to α:

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \psi_\tau(y - \alpha) dF(y) \\ &= (\tau - 1) \int_{-\infty}^{\alpha} dF(y) + \tau \int_{\alpha}^{\infty} dF(y) \\ &= (\tau - 1)F(\alpha) + \tau(1 - F(\alpha)) \end{aligned}$$

implying $\tau = F(\alpha)$ and thus $\hat{\alpha} = F^{-1}(\tau)$.

Sample Quantiles as Optimizers

For sample quantiles replace F by $\hat{F}$, the empirical distribution function. The objective function becomes a polyhedral convex function whose derivative is monotone decreasing, in effect the gradient simply counts observations above and below and weights the sums by τ and $\tau - 1$.



Conditional Quantiles: The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \text{argmin}_m \mathbb{E}(Y - m)^2$$

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

$$\mu(x) = \text{argmin}_m \mathbb{E}_{Y|X=x}(Y - m(X))^2.$$

Similarly, the unconditional τth quantile solves

$$\alpha_\tau = \text{argmin}_a \mathbb{E}\rho_\tau(Y - a)$$

and the conditional τth quantile solves

$$\alpha_\tau(x) = \text{argmin}_a \mathbb{E}_{Y|X=x}\rho_\tau(Y - a(X))$$

Computation of Linear Regression Quantiles

Primal Formulation as a linear program, split the residual vector into positive and negative parts and sum with appropriate weights:

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

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.
Let $h \in \mathcal{H}$ index p-element subsets of $\{1, 2, ..., n\}$ then primal solutions take the form:

$$\hat{\beta} = \hat{\beta}(h) = X(h)^{-1} y(h)$$

Least Squares from the p-subset Perspective

Exact fits to p observations:

$$\hat{\beta} = \hat{\beta}(h) = X(h)^{-1} y(h)$$

OLS is a weighted average of these $\hat{\beta}(h)$'s:

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y = \sum_{h \in \mathcal{H}} w(h) \hat{\beta}(h),$$

$$w(h) = |X(h)|^2 / \sum_{h \in \mathcal{H}} |X(h)|^2$$

The determinants $|X(h)|$ are the (signed) volumes of the parallelipipeds formed by the columns of the the matrices $X(h)$. In the simplest bivariate case, we have,

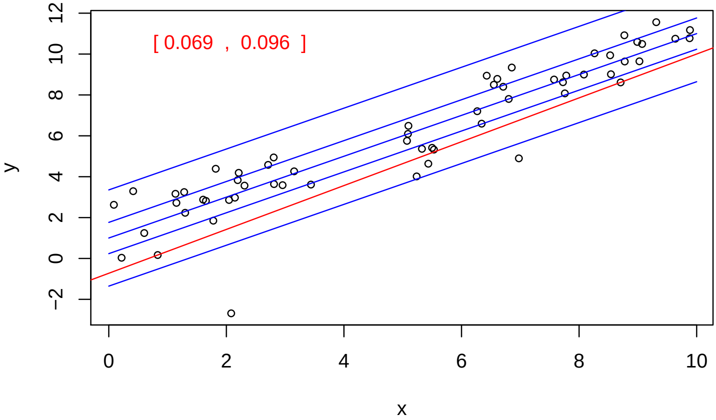
$$|X(h)|^2 = \begin{vmatrix} 1 & x_i \\ 1 & x_j \end{vmatrix}^2 = (x_j - x_i)^2$$

so pairs of observations that are far apart are given more weight.

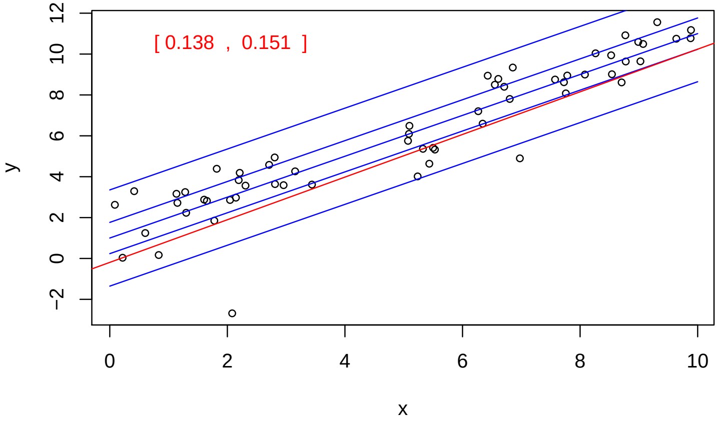
Quantile Regression: The Movie

- Bivariate linear model with iid Student t errors
- Conditional quantile functions are parallel in blue
- 100 observations indicated in blue
- Fitted quantile regression lines in red.
- Intervals for $\tau \in (0, 1)$ for which the solution is optimal.

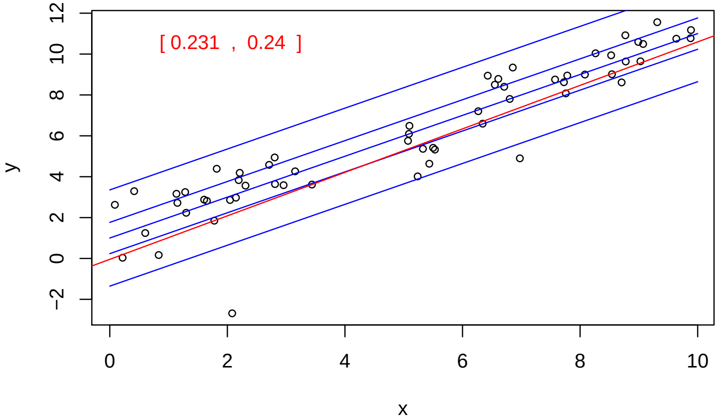
Quantile Regression in the iid Error Model



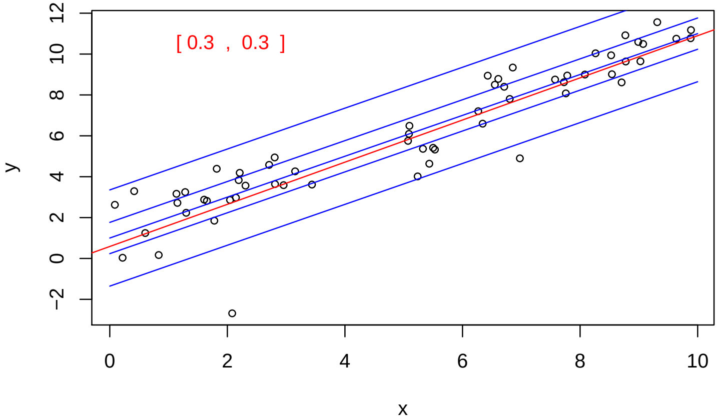
Quantile Regression in the iid Error Model



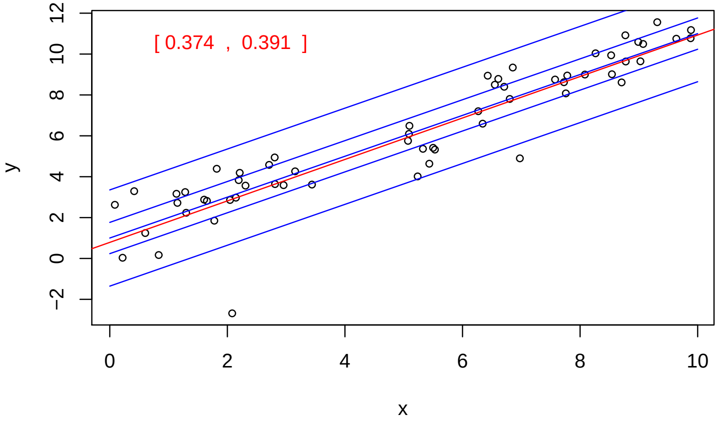
Quantile Regression in the iid Error Model



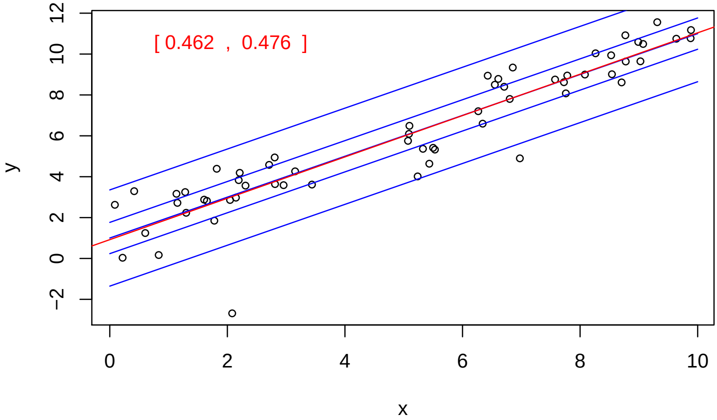
Quantile Regression in the iid Error Model



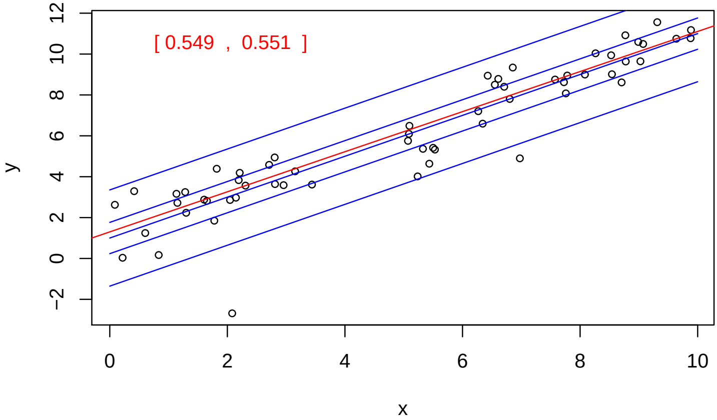
Quantile Regression in the iid Error Model



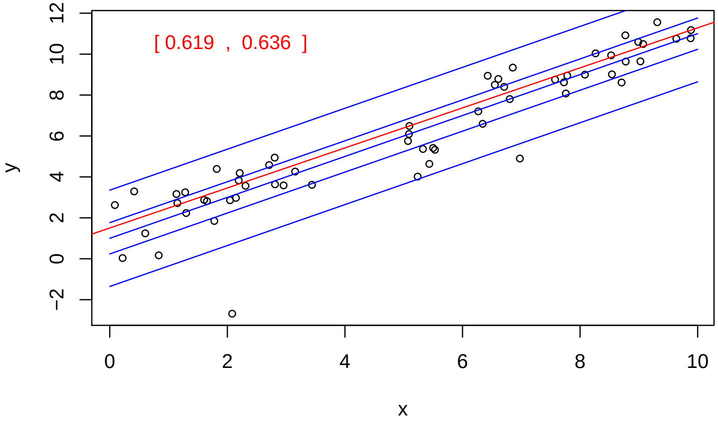
Quantile Regression in the iid Error Model



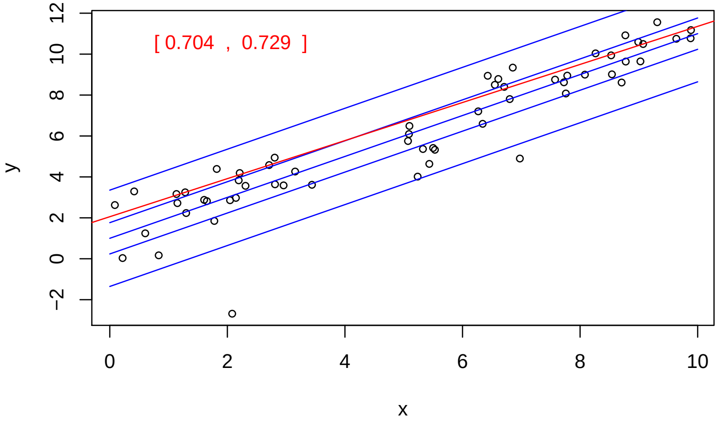
Quantile Regression in the iid Error Model



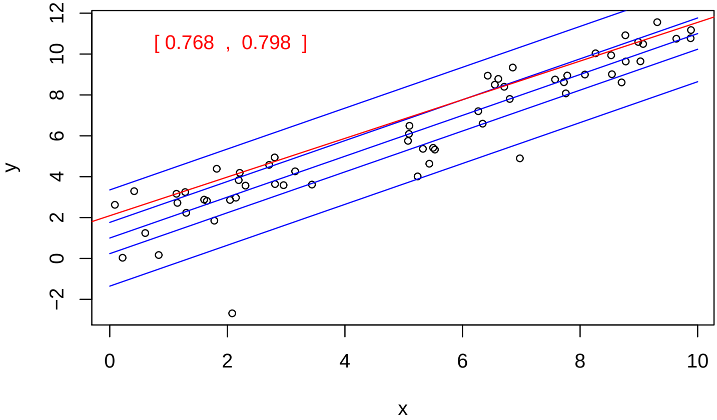
Quantile Regression in the iid Error Model



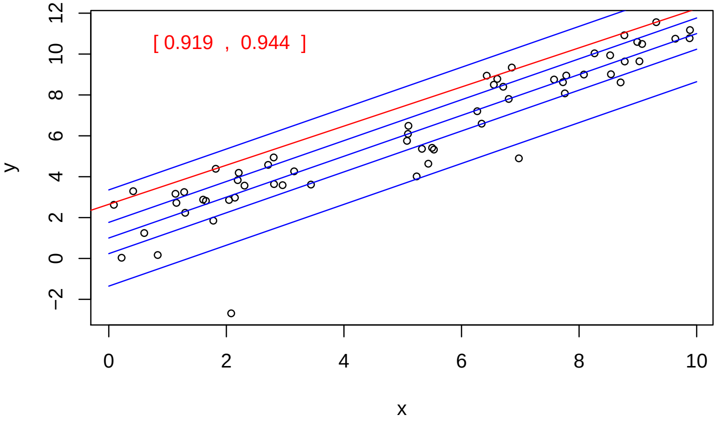
Quantile Regression in the iid Error Model



Quantile Regression in the iid Error Model



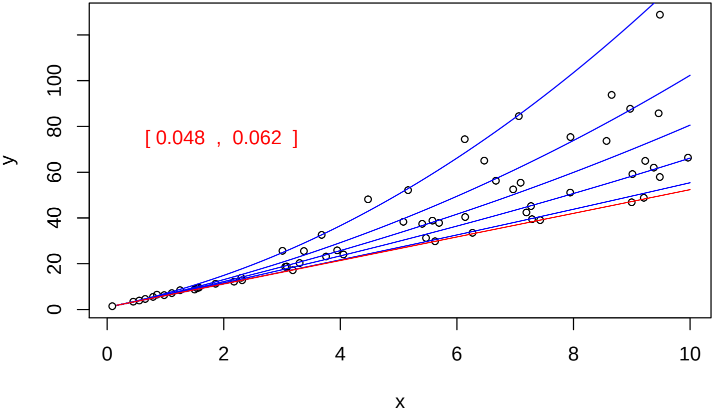
Quantile Regression in the iid Error Model



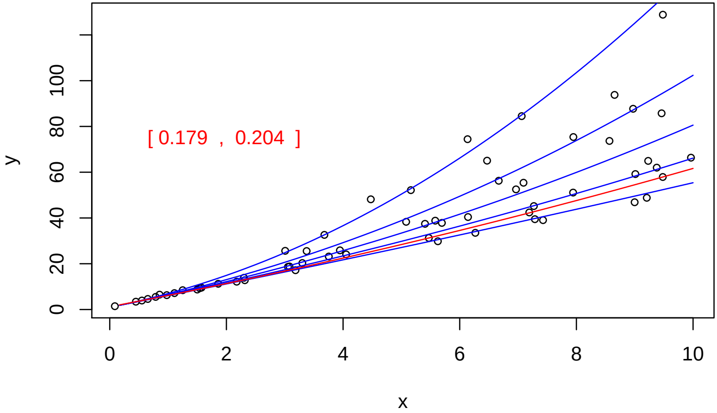
Virtual Quantile Regression II

- Bivariate quadratic model with Heteroscedastic χ^2 errors
- Conditional quantile functions drawn in blue
- 100 observations indicated in blue
- Fitted quadratic quantile regression lines in red
- Intervals of optimality for $\tau \in (0, 1)$.

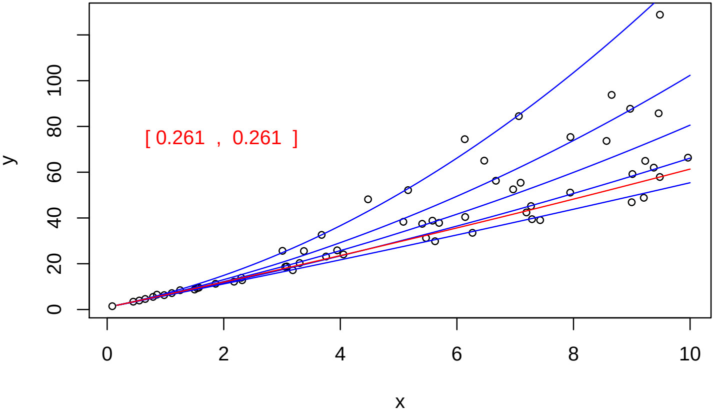
Quantile Regression in the Heteroscedastic Error Model



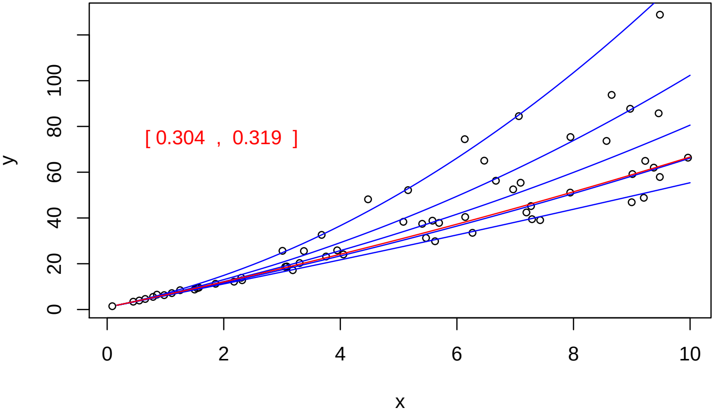
Quantile Regression in the Heteroscedastic Error Model



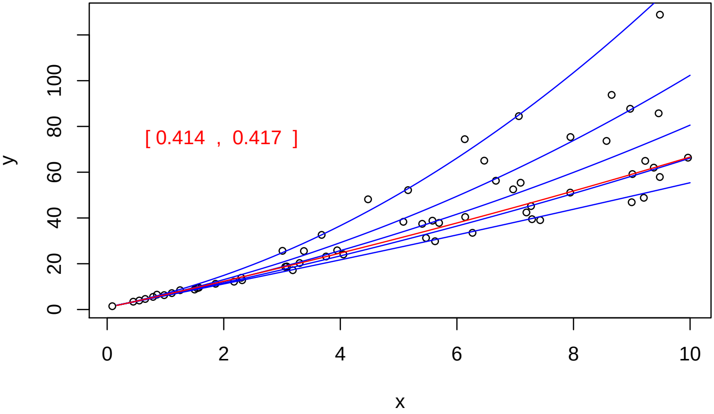
Quantile Regression in the Heteroscedastic Error Model



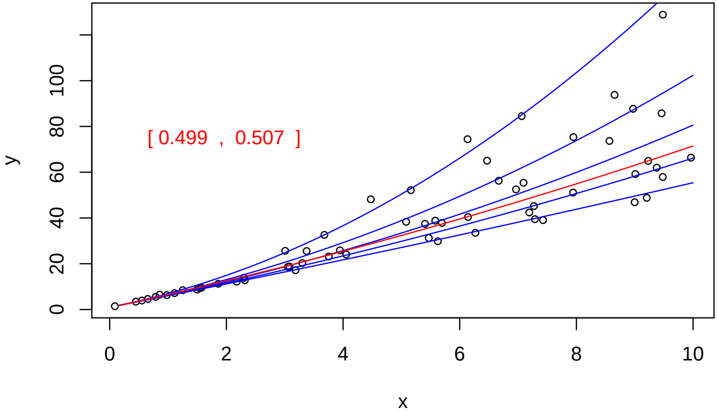
Quantile Regression in the Heteroscedastic Error Model



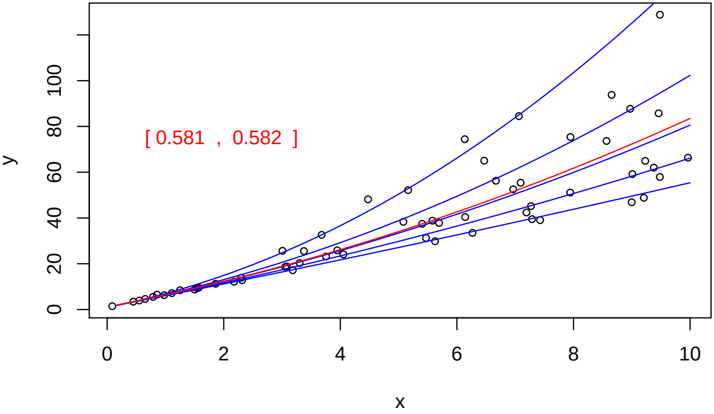
Quantile Regression in the Heteroscedastic Error Model



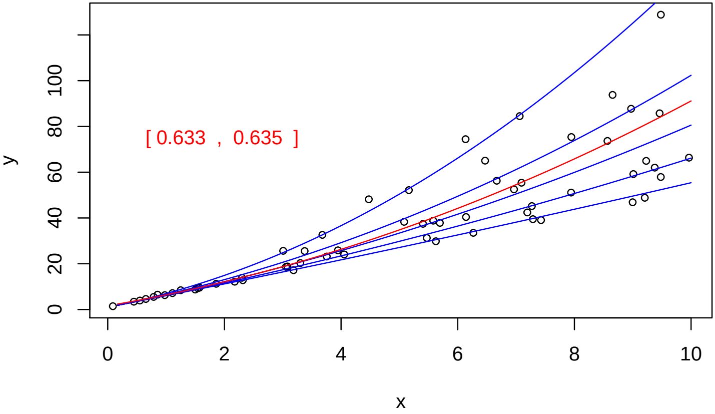
Quantile Regression in the Heteroscedastic Error Model



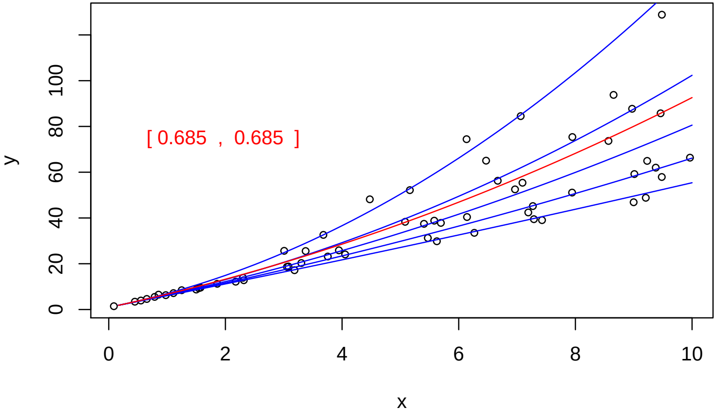
Quantile Regression in the Heteroscedastic Error Model



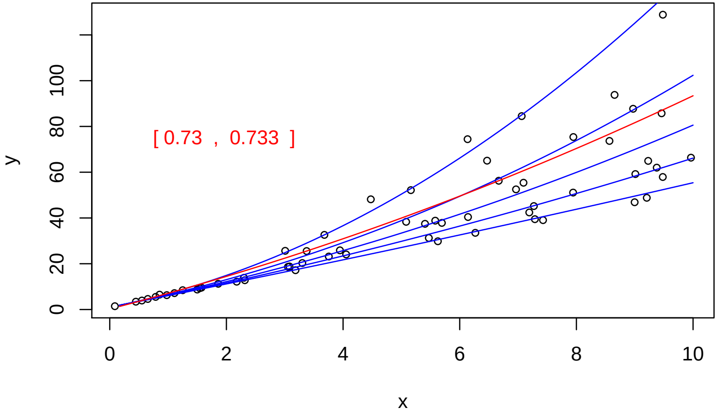
Quantile Regression in the Heteroscedastic Error Model



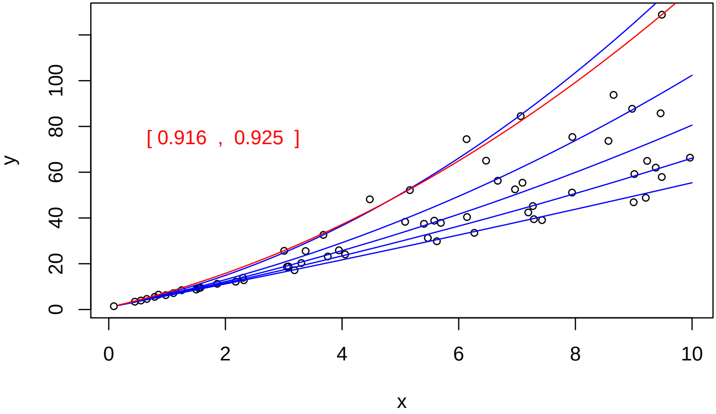
Quantile Regression in the Heteroscedastic Error Model



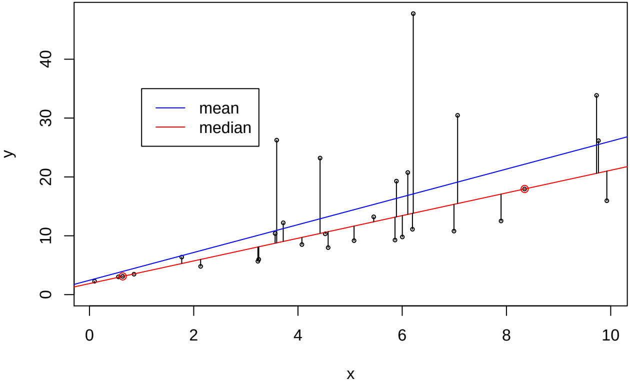
Quantile Regression in the Heteroscedastic Error Model



Quantile Regression in the Heteroscedastic Error Model



Conditional Means vs. Medians



Minimizing absolute errors for median regression can yield something quite different from the least squares fit for mean regression.

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 $|A| \neq 0$, $\hat{\beta}(\tau; y, AX) = A^{-1}\hat{\beta}(\tau; yX)$
- Robustness: For any diagonal matrix D with nonnegative elements. $\hat{\beta}(\tau; y, X) = \hat{\beta}(\tau, y + D\hat{u}, X)$

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, generally,

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

Examples:
$h(y) = \min\{0, y\}$, Powell's (1985) censored regression estimator.
$h(y) = \text{sgn}\{y\}$ Rosenblatt's (1957) perceptron, Manski's (1975) maximum score estimator. estimator.

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.*

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

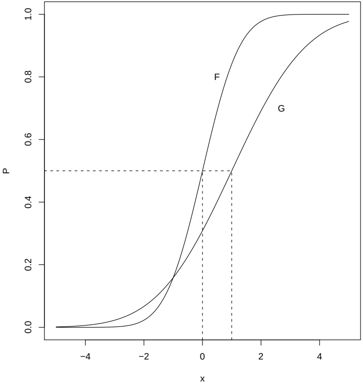
Then $\Delta(x)$ is uniquely defined as

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

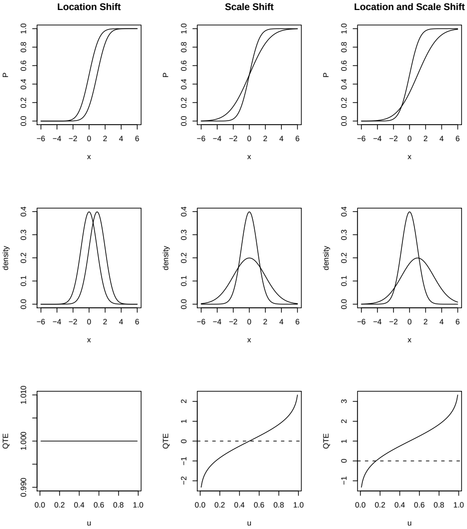
This is the essence of the conventional QQ-plot. Changing variables so $\tau = F(x)$ we have the quantile treatment effect (QTE):

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

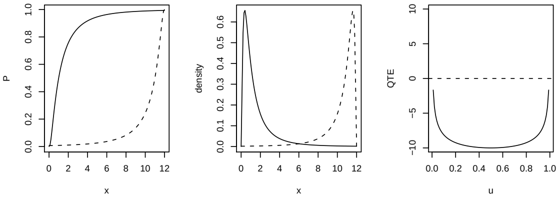
Lehmann-Doksum QTE



Lehmann-Doksum QTE



An Asymmetric Example



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

The Erotic is Unidentified

The Lehmann QTE characterizes the difference in the marginal distributions, F and G, but it cannot reveal anything about the joint distribution, H. The copula function, Schweizer and Wolf (1981), Genest and McKay, (1986),

$$\varphi(u, v) = H(F^{-1}(u), G^{-1}(v)),$$

is *not* identified. Lehmann's formulation *assumes* that the treatment leaves the ranks of subjects invariant. If a subject was going to be the median control subject, then he will also be the median treatment subject. This is an inherent limitation of the Neymann-Rubin potential outcomes framework.

QTE via Quantile Regression

The Lehmann QTE is naturally estimable by

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

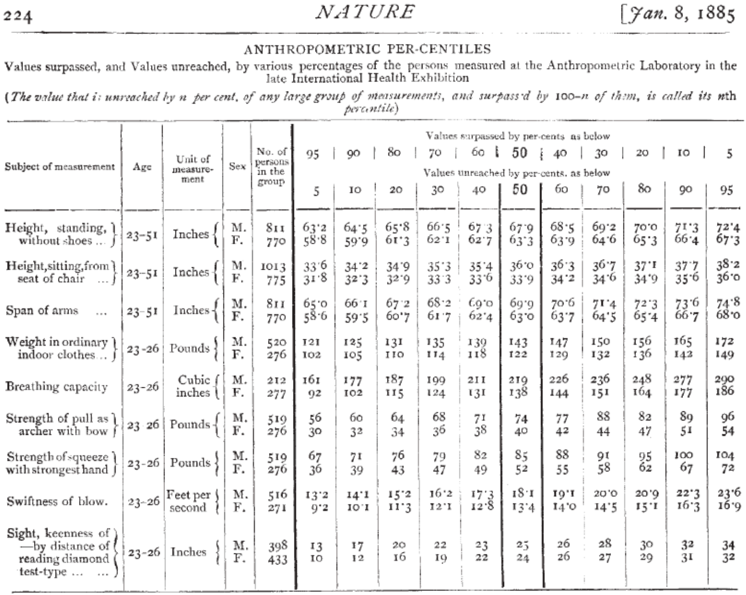
where $\hat{G}_n$ and $\hat{F}_m$ denote the empirical distribution functions of the treatment and control observations, Consider the quantile regression model

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

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,

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

Francis Galton's (1885) Anthropometric Quantiles



224 *NATURE* [*Jan.* 8, 1885

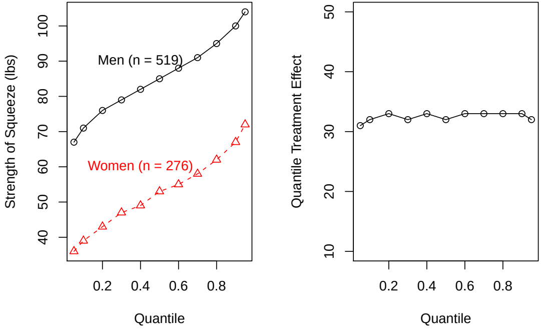
ANTHROPOMETRIC PER-CENTILES

Values surpassed, and Values unreached, by various percentages of the persons measured at the Anthropometric Laboratory in the late International Health Exhibition

(*The value that is unreached by n per cent. of any large group of measurements, and surpass'd by* 100–*n of them, is called its n*th *percentile*)

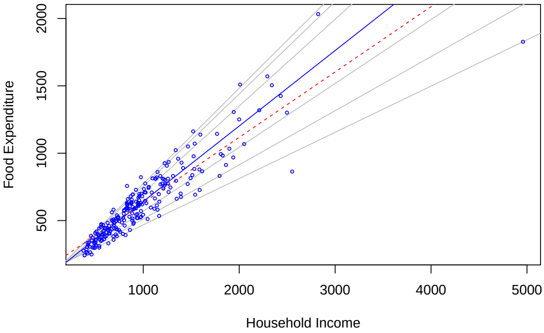
Subject of measurement	Age	Unit of measurement	Sex	No. of persons in the group	Values surpassed by per-cents as below										
					95	90	80	70	60	50	40	30	20	10	5
					Values unreached by per-cents. as below										
					5	10	20	30	40	50	60	70	80	90	95
Height, standing, without shoes ...	23–51	Inches	M. F.	811 770	63·2 58·8	64·5 59·9	65·8 61·3	66·5 62·1	67·3 62·7	67·9 63·3	68·5 63·9	69·2 64·6	70·0 65·3	71·3 66·4	72·4 67·3
Height, sitting, from seat of chair ...	23–51	Inches	M. F.	1013 775	33·6 31·8	34·2 32·3	34·9 32·9	35·3 33·3	35·4 33·6	36·0 33·9	36·3 34·2	36·7 34·6	37·1 34·9	37·7 35·6	38·2 36·0
Span of arms ...	23–51	Inches	M. F.	811 770	65·0 58·6	66·1 59·5	67·2 60·7	68·2 61·7	69·0 62·4	69·9 63·0	70·6 63·7	71·4 64·5	72·3 65·4	73·6 66·7	74·8 68·0
Weight in ordinary indoor clothes ...	23–26	Pounds	M. F.	520 276	121 102	125 105	131 110	135 114	139 118	143 122	147 129	150 132	156 136	165 142	172 149
Breathing capacity	23–26	Cubic inches	M. F.	212 277	161 92	177 102	187 115	199 124	211 131	219 138	226 144	236 151	248 164	277 177	290 186
Strength of pull as archer with bow	23 26	Pounds	M. F.	519 276	56 30	60 32	64 34	68 36	71 38	74 40	77 42	88 44	82 47	89 51	96 54
Strength of squeeze with strongest hand	23–26	Pounds	M. F.	519 276	67 36	71 39	76 43	79 47	82 49	85 52	88 55	91 58	95 62	100 67	104 72
Swiftness of blow.	23–26	Feet per second	M. F.	516 271	13·2 9·2	14·1 10·1	15·2 11·3	16·2 12·1	17·3 12·8	18·1 13·4	19·1 14·0	20·0 14·5	20·9 15·1	22·3 16·3	23·6 16·9
Sight, keenness of—by distance of reading diamond test-type	23–26	Inches	M. F.	398 433	13 10	17 12	20 16	22 19	23 22	25 24	26 26	28 27	30 29	32 31	34 32

Quantile Treatment Effects: Strength of Squeeze



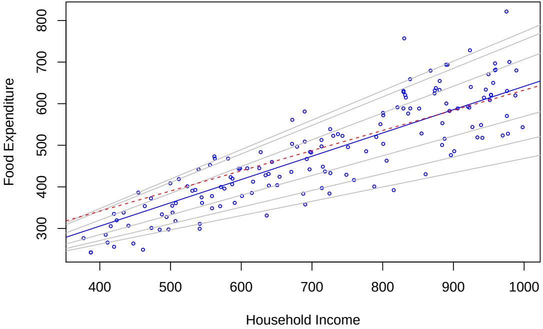
"Very powerful women exist, but happily perhaps for the repose of the other sex, such gifted women are rare."

Engel's Food Expenditure Data



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 blue solid line; the least squares estimate of the conditional mean function is indicated by the red dashed line.

Engel's Food Expenditure Data

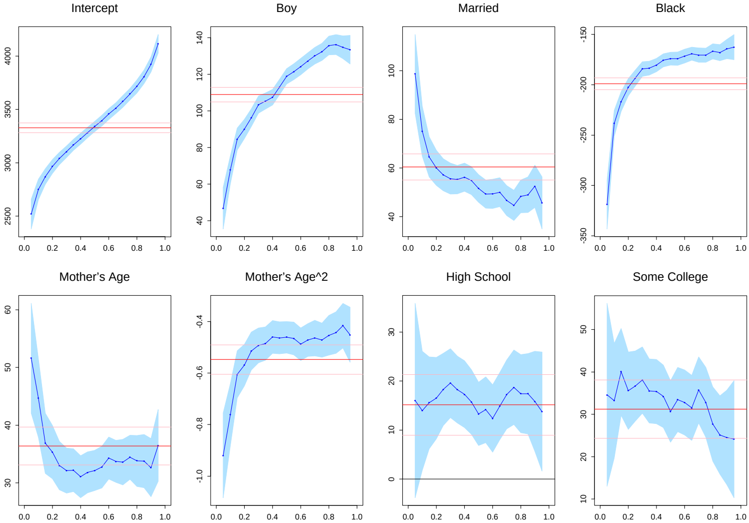


Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$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

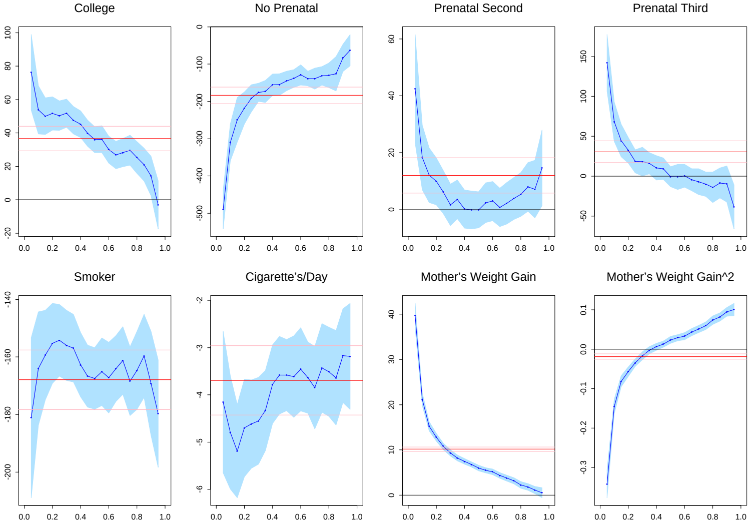
A Model of Infant Birthweight

- 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

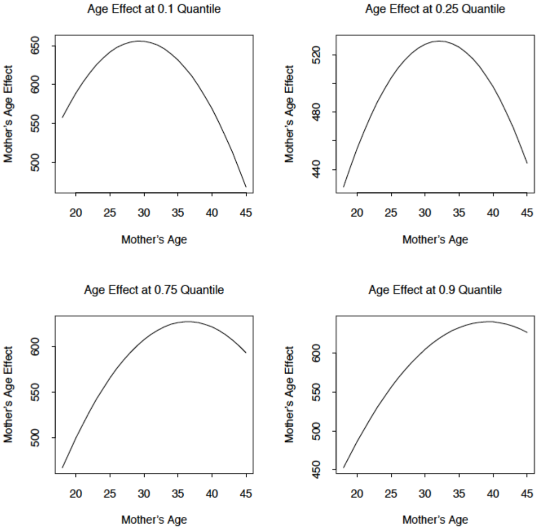
Quantile Regression Birthweight Model I



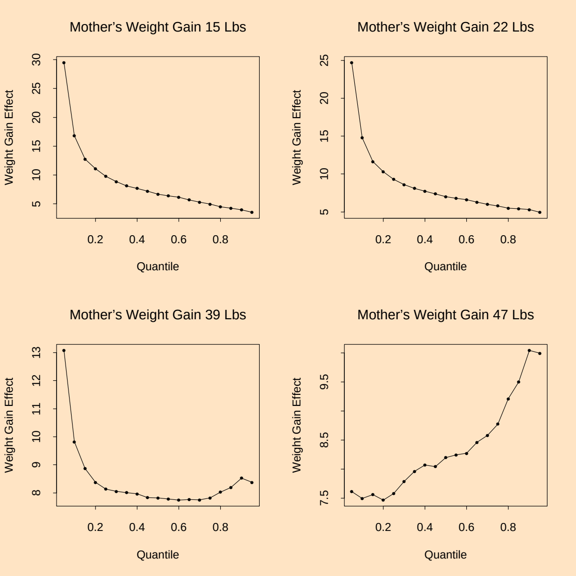
Quantile Regression Birthweight Model II



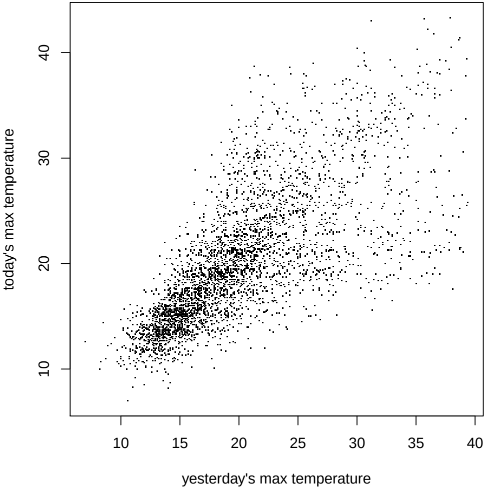
Marginal Effect of Mother’s Age



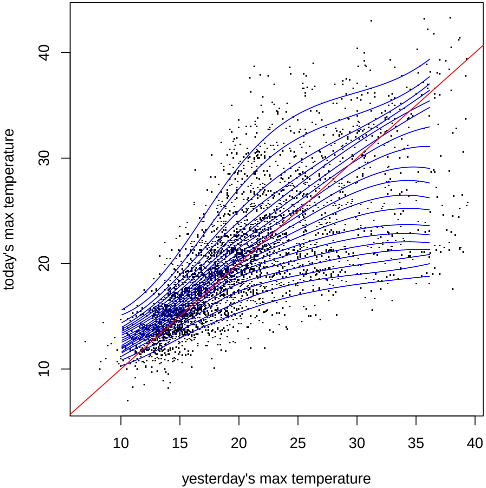
Marginal Effect of Mother's Weight Gain



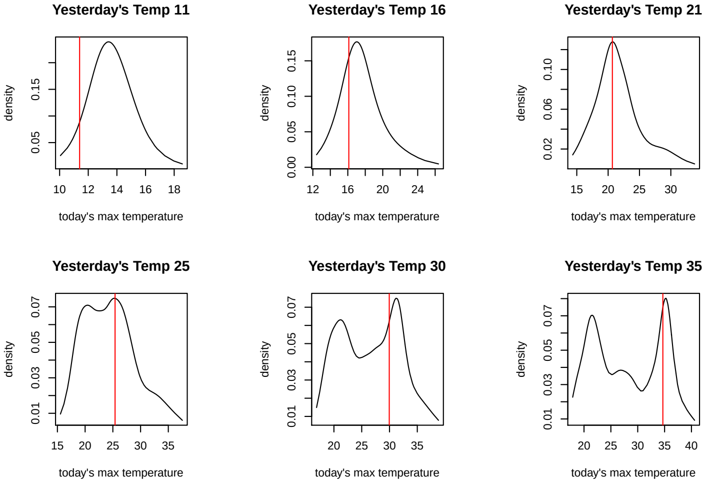
Daily Temperature in Melbourne: AR(1) Scatterplot



Daily Temperature in Melbourne: Nonlinear QAR(1) Fit



Conditional Densities of Melbourne Daily Temperature



Review of Lecture 1

Least squares meethods of estimating conditional mean functions

- were developed for, and
- promote the view that,

$$\text{Response} = \text{Signal} + \text{iid Measurement Error}$$

In fact the world is rarely this simple.