Quantile Regression
A JSM Short Course

Roger Koenker & Judy Huixia Wang

University of Illinois, Urbana-Champaign & George Washington University

August 02, 2014

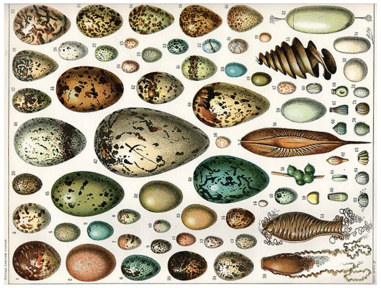
Overview of the Lectures

1. Lecture 1. The Basics: What, Why and How?
2. Lecture 2. Inference for Quantile Regression
3. Lecture 3. Computation and Examples
4. Lecture 4. Censored QR and Survival Analysis
5. Lecture 5. Nonparametric Quantile Regression
6. Lecture 6. Bayesian Quantile Regression
7. Lecture 7. Quantile Regression for Longitudinal Data

Lecture 1. The Basics: What, Why and How?

- Univariate Quantiles
- Scatterplot Smoothing
- Equivariance Properties
- Quantile Treatment Effects
- Three Empirical Examples

The Middle Sized Egg



Volume of the eggs can be measured by the amount of water they displace (Archimedes' Eureka!) 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!

Stem and Leaf Plot: Tukey's EDA I

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

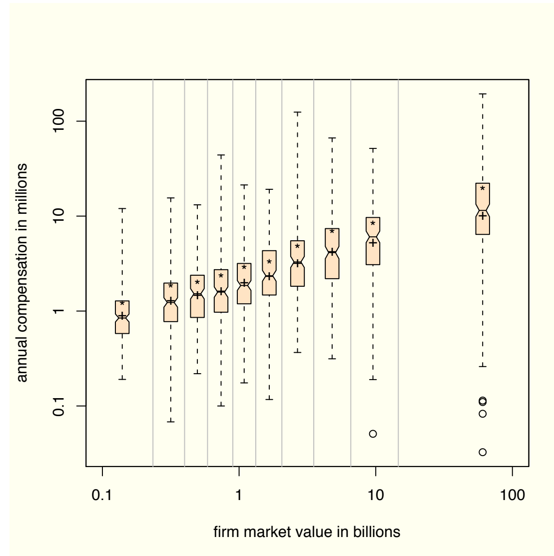
```
> x <- rchisq(100,5) # 100 Chi-squared(5)
> quantile(x) # Tukey's Five Number Summary
        0%        25%        50%        75%       100%
 0.9042396  2.7662230  4.2948642  6.2867588 16.5818573

> stem(x)

  The decimal point is at the |

   0 | 92356668
   2 | 0011112444456677788899990111222455666
   4 | 01223334666678901125567889
   6 | 023344667802888
   8 | 556691
  10 | 7
  12 | 159
  14 | 06
  16 | 6
```


Boxplot of CEO Pay: Tukey's EDA II



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.

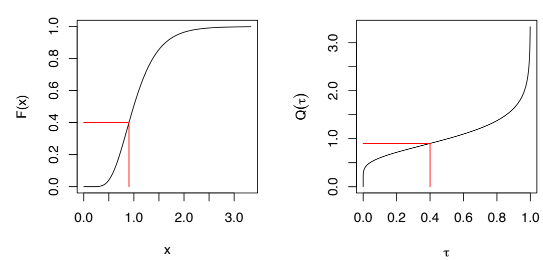
Mosteller and Tukey (1977)

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.

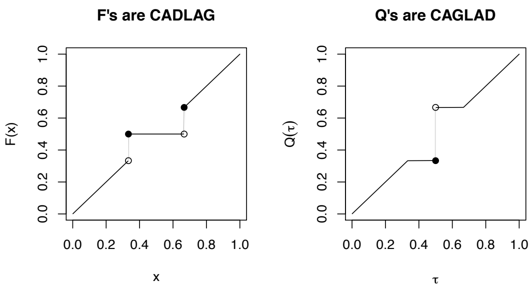


Univariate Quantiles

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

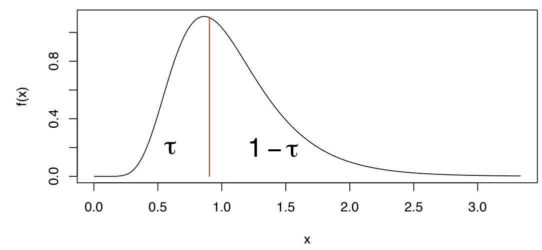
$$Q_X(\tau) = F_X^{-1}(\tau) = \inf\{x \mid F(x) \geqslant \tau\}.$$

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



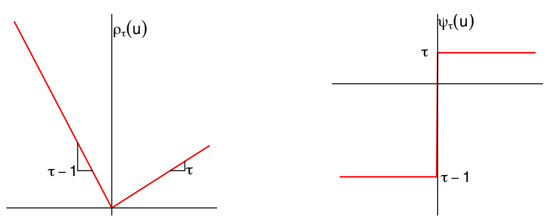
Univariate Quantiles

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:

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

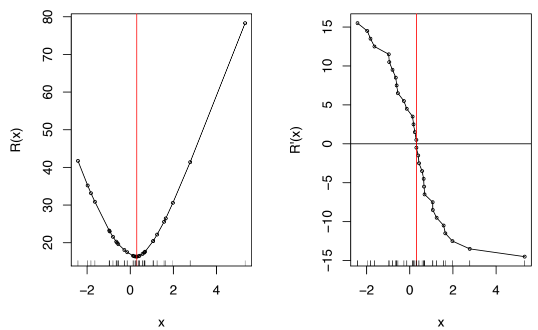
Proof: Let $\psi_\tau(u) = \rho'_\tau(u)$, so differentiating wrt to α:

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

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.

Conditional vs Marginal Quantiles

Interpretation of QR results must be careful to distinguish conditional and marginal covariate effects:

- QR estimates covariate effects on conditional quantiles,
- How do changes in covariates impact conditional quantiles of the response,
- Such effects may depend crucially on what the other conditioning covariates are,
- All of this is familiar from classical mean (least-squares) regression,
- But, perhaps, too easily overlooked.

Effects on marginal quantiles may be estimated by integrating with respect to the marginal distribution of the covariates, Machado and Mata (2001), Chernozhukov et al. (2013).

What is a College Degree Worth?

$$Y_i = \alpha_i + \beta_i + \gamma_i,$$

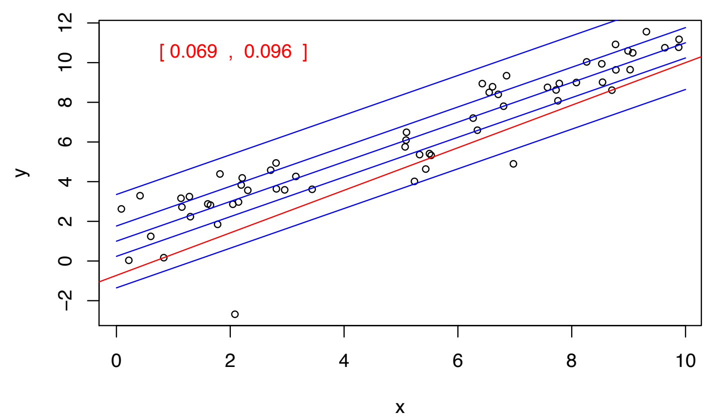
- Y_i Discounted Career Earnings
- α_i Ability Component
- β_i College Component
- γ_i Major Component

Conditional Quantile Interpretation: Among individuals of a given α and γ, what income level, β, is such that proportion τ are below β, and $(1-\tau)$ are above.
Conditional Distribution Interpretation: Given a fixed income level, y, how do changes in ability, college completion status, and choice of major influence the probability of exceeding that level.

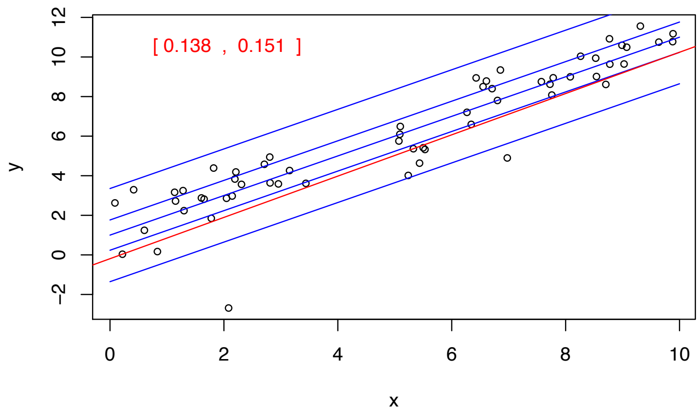
Quantile Regression: The Movie

- 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.

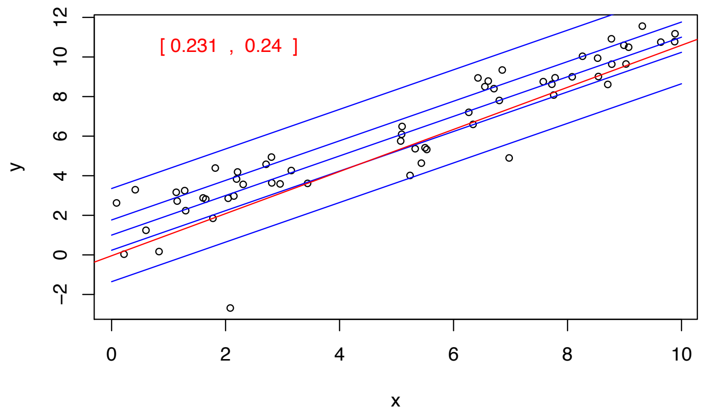
Quantile Regression in the iid Error Model



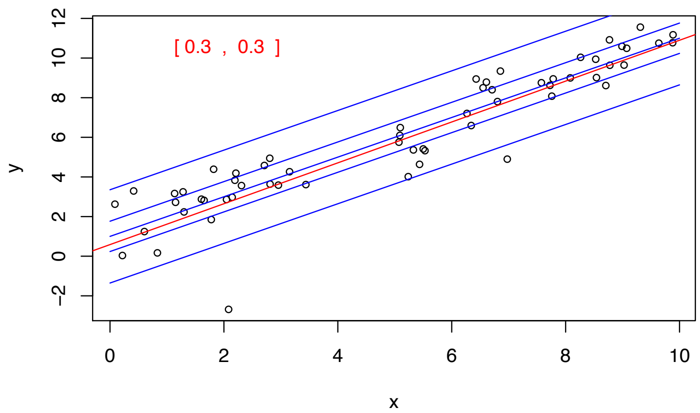
Quantile Regression in the iid Error Model



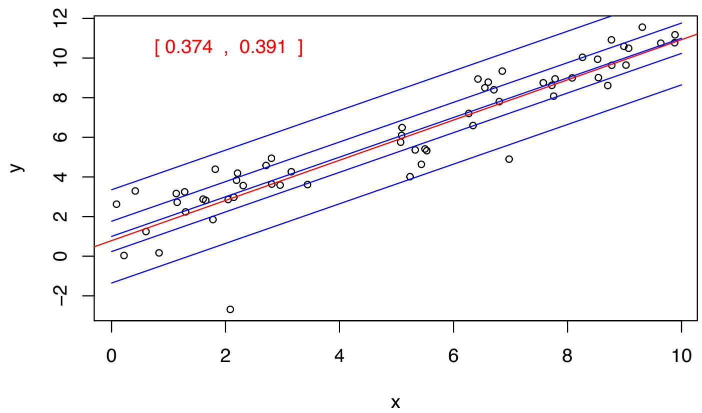
Quantile Regression in the iid Error Model



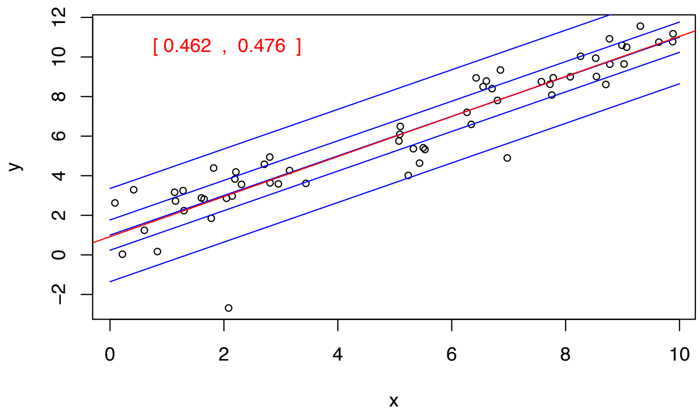
Quantile Regression in the iid Error Model



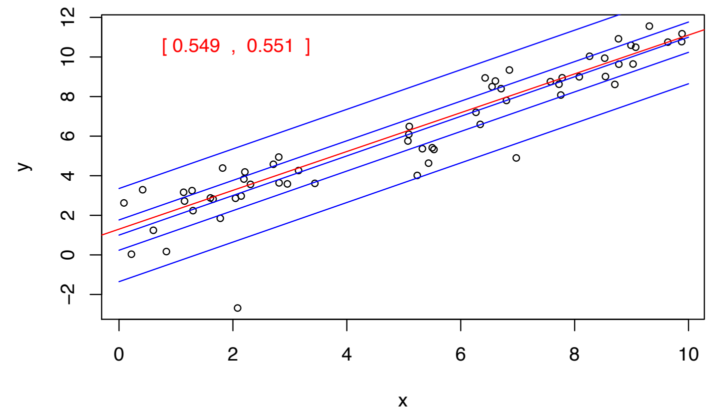
Quantile Regression in the iid Error Model



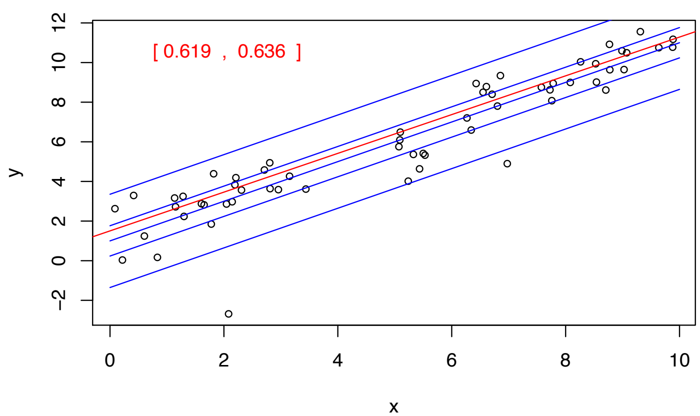
Quantile Regression in the iid Error Model



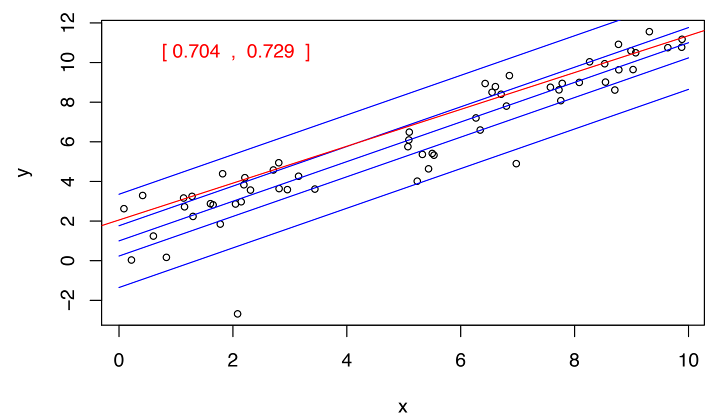
Quantile Regression in the iid Error Model



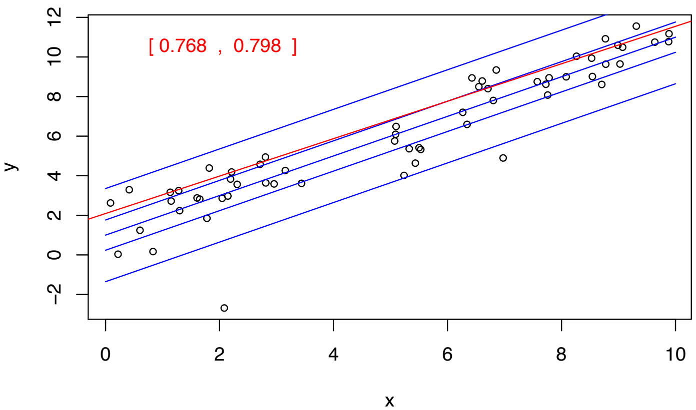
Quantile Regression in the iid Error Model



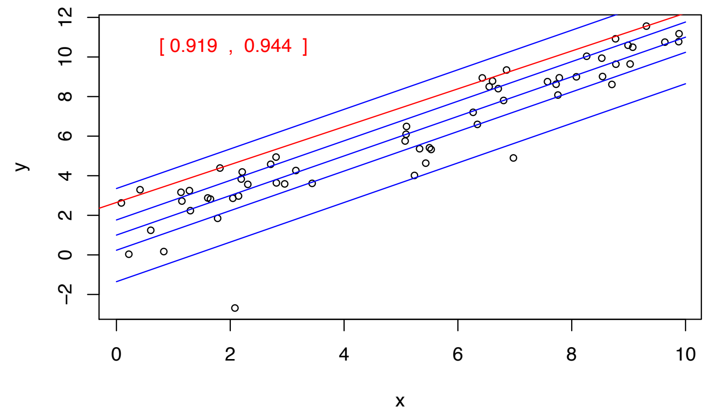
Quantile Regression in the iid Error Model



Quantile Regression in the iid Error Model



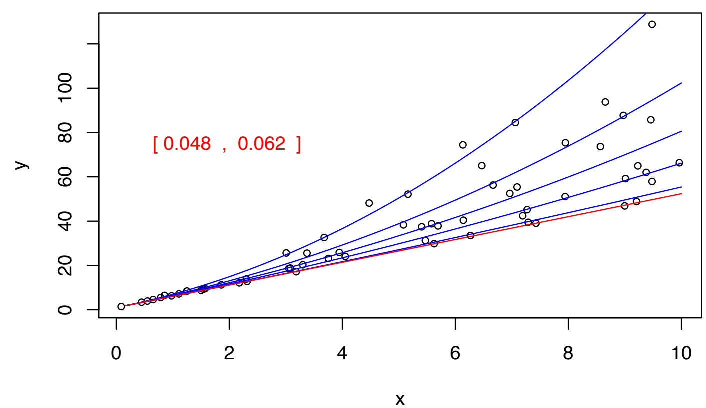
Quantile Regression in the iid Error Model



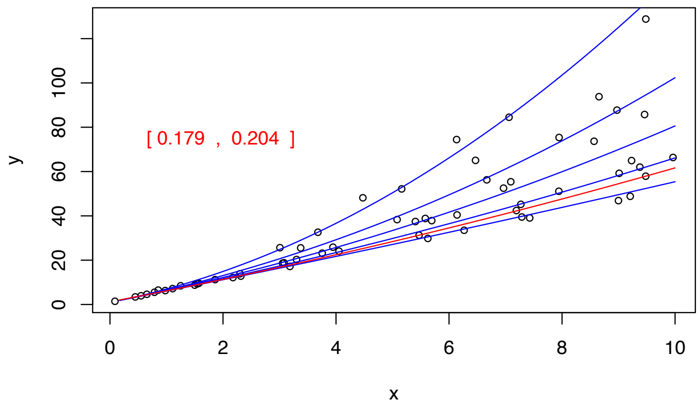
Virtual Quantile Regression II

- 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 $\tau \in (0,1)$.

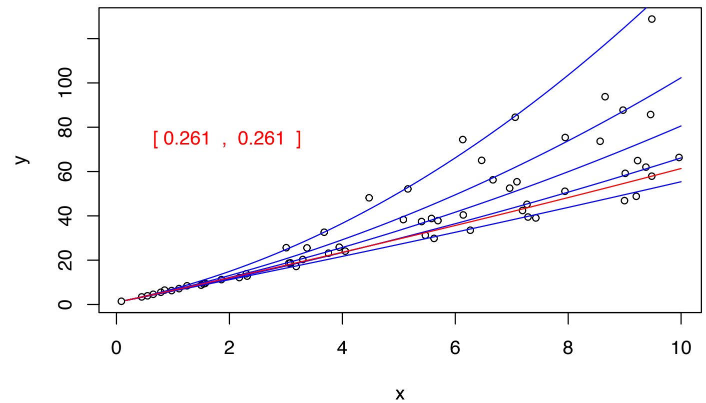
Quantile Regression in the Heteroscedastic Error Model



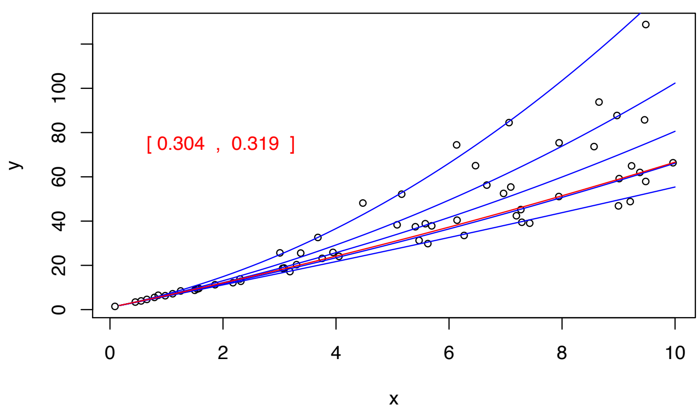
Quantile Regression in the Heteroscedastic Error Model



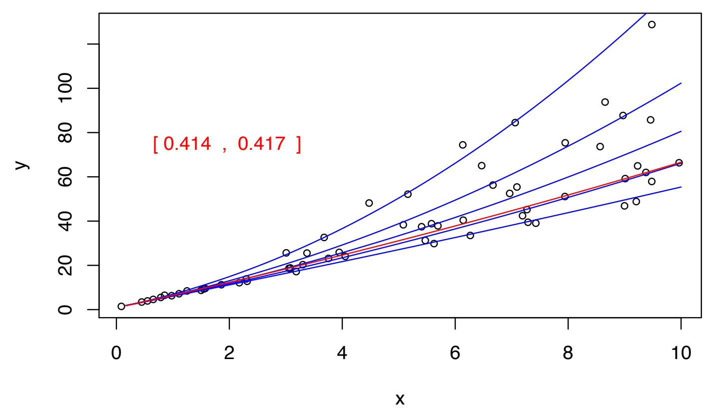
Quantile Regression in the Heteroscedastic Error Model



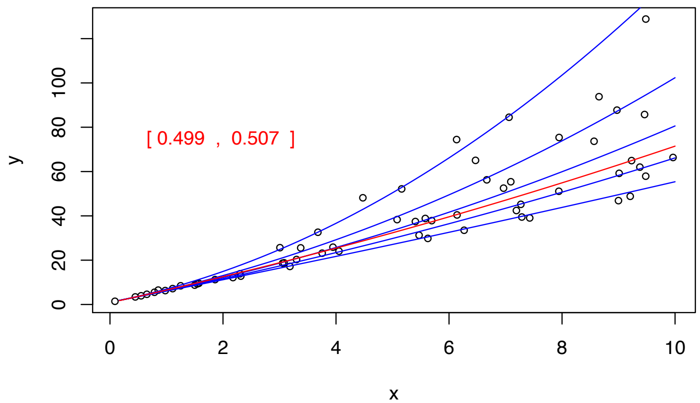
Quantile Regression in the Heteroscedastic Error Model



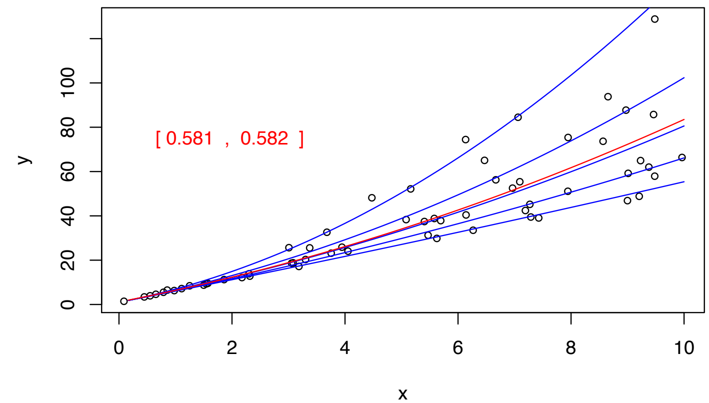
Quantile Regression in the Heteroscedastic Error Model



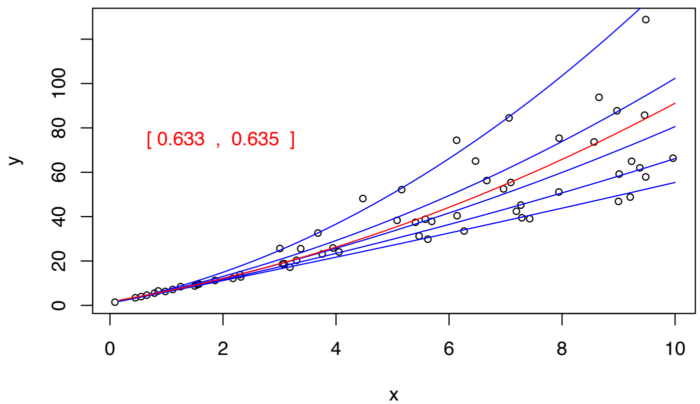
Quantile Regression in the Heteroscedastic Error Model



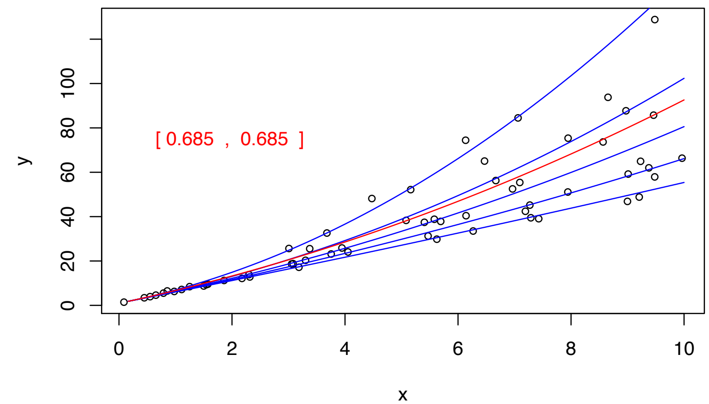
Quantile Regression in the Heteroscedastic Error Model



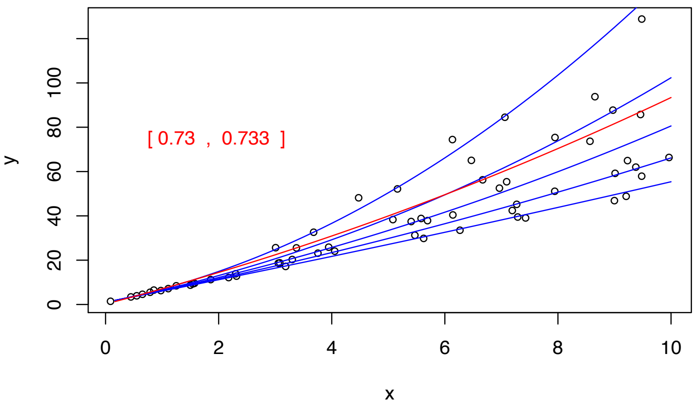
Quantile Regression in the Heteroscedastic Error Model



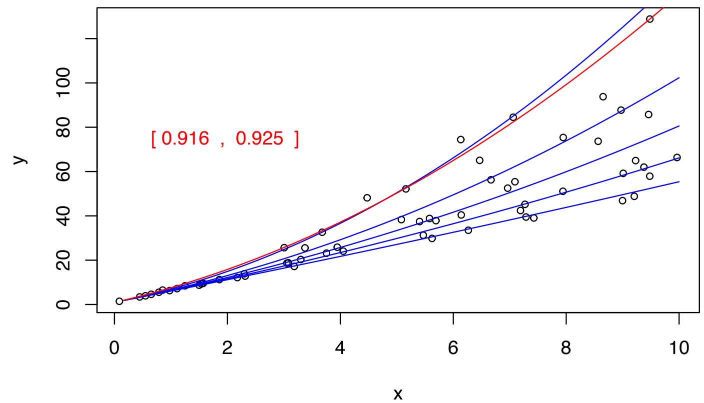
Quantile Regression in the Heteroscedastic Error Model



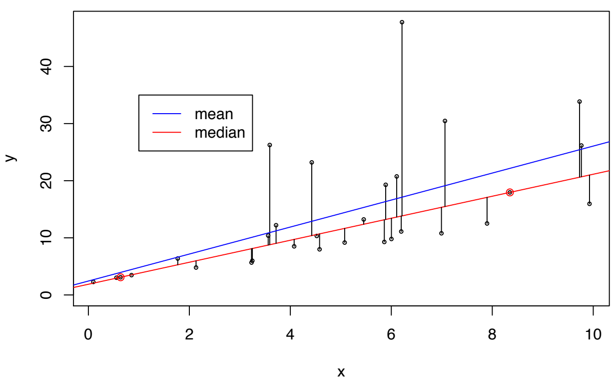
Quantile Regression in the Heteroscedastic Error Model



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

$$\hat{\beta}(\tau; y, X) = \text{argmin}_b \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b)$$

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

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:
$h(y) = \min\{0, y\}$, censored regression estimator of Powell (1986) .
$h(y) = \text{sgn}\{y\}$, perceptron of Rosenblatt (1958), maximum score estimator of Manski (1975).

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.*

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

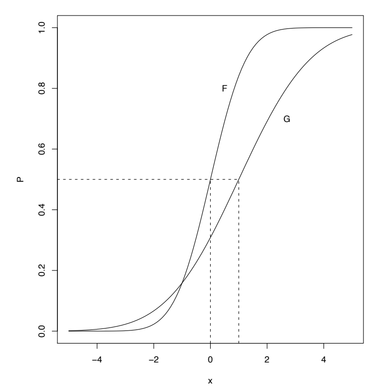
Then $\Delta(x)$ is uniquely defined as

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

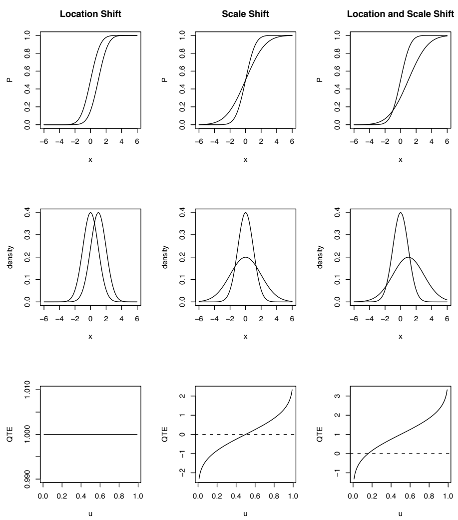
This is the essence of the conventional QQ-plot. Changing variables so $\tau = F(x)$ we have the quantile treatment effect (QTE):

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

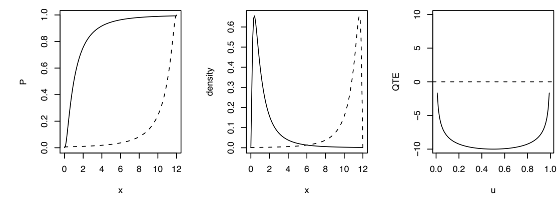
Lehmann-Doksum QTE



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 observation. 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,

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

Galton's (1885) Anthropometric Quantiles

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

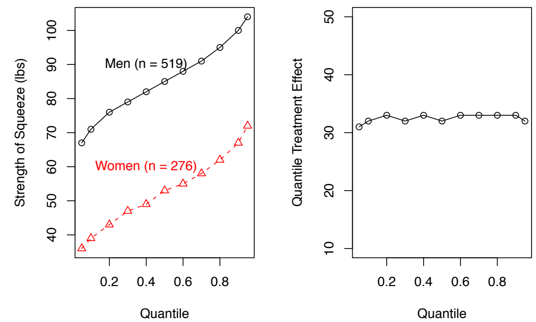
ANTHROPOMETRIC PER-CENTILES

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 surpassed by 100−n of them, is called its nth percentile*)

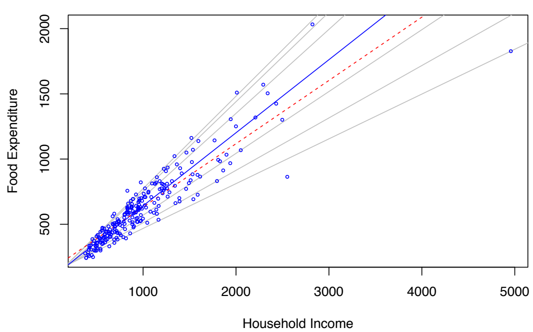
Subject of measurement	Age	Unit of measurement	Sex	No. of persons in the group	Values surpassed by per-cents as below: 95	90	80	70	60	50	40	30	20	10	5
					Values unreached by per-cents, as below: 5	10	20	30	40	50	60	70	80	90	95
Height, standing, without shoes ...	23–51	Inches	M.	811	63·2	64·5	65·8	66·5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

QTE: Strength of Squeeze



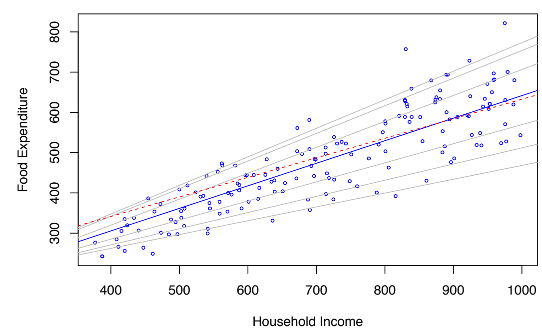
"Very powerful women exist, but happily perhaps for the repose of the other sex, such gifted women are rare."

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

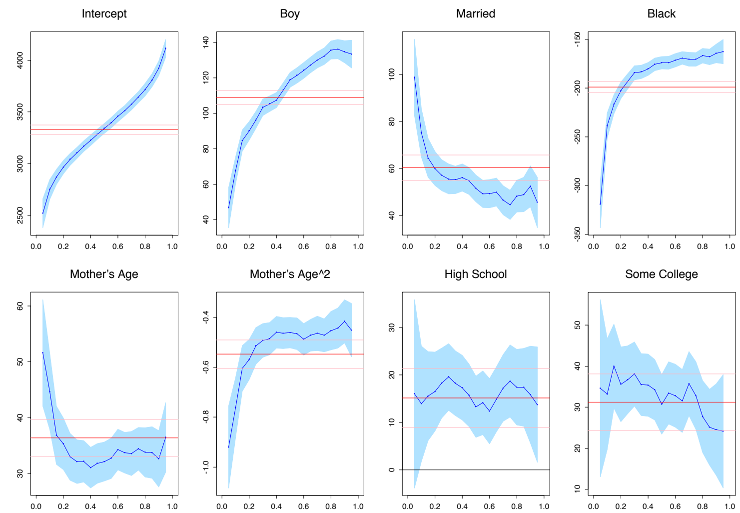


Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$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

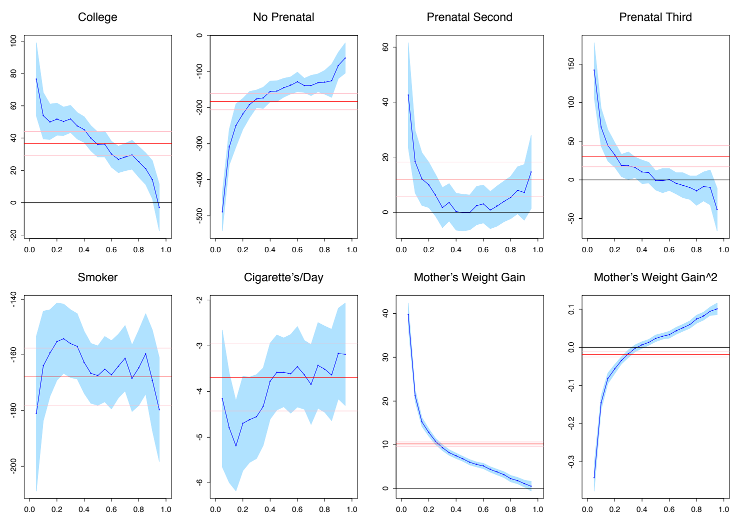
A Model of Infant Birthweight

- Reference: Abrevaya (2001), Koenker and Hallock (2000).
- 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

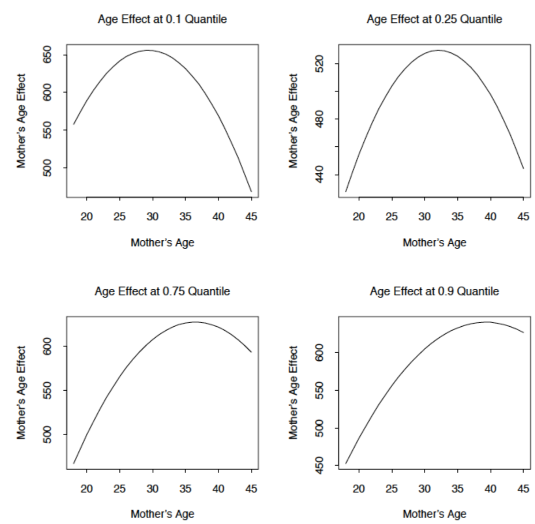
Quantile Regression Birthweight Model I



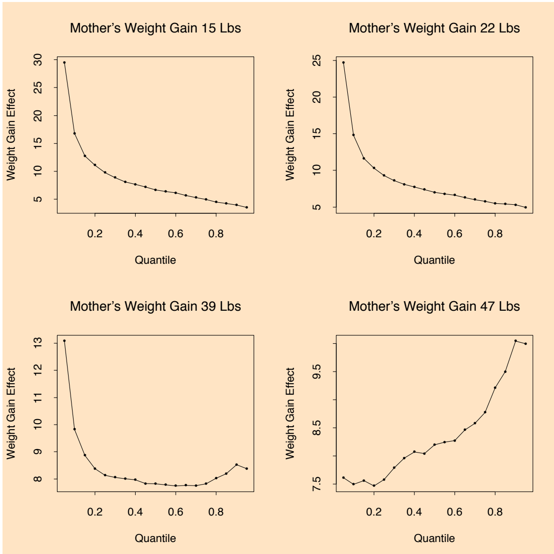
Quantile Regression Birthweight Model II



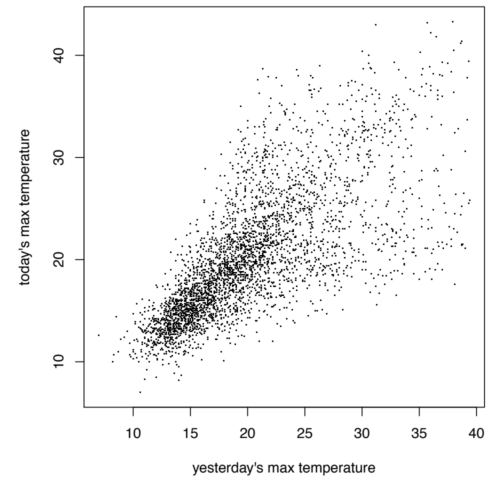
Marginal Effect of Mother's Age



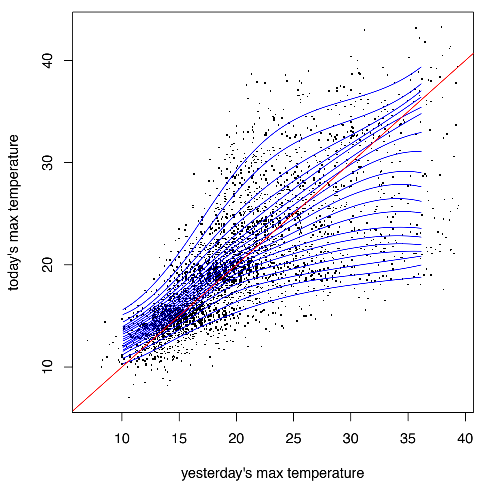
Marginal Effect of Mother's Weight Gain



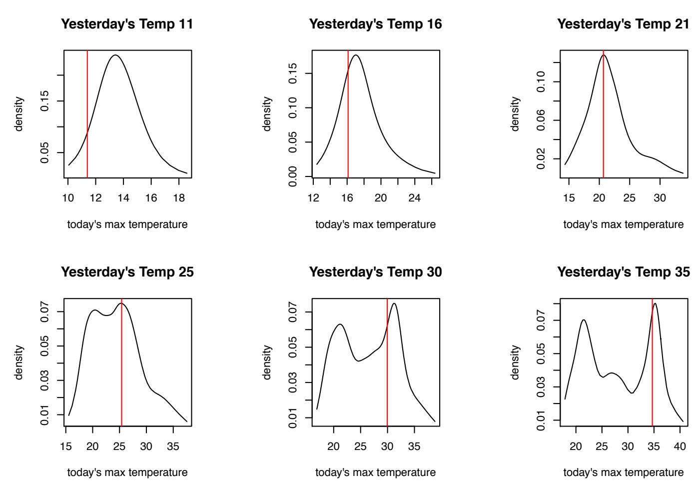
Daily Temperature in Melbourne



Nonparametric QAR(1) Fit



Conditional Densities of Daily Temperature



Review of Lecture 1

Least squares methods 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.
More flexible models can be formulated that allow one to estimate several conditional quantile functions of the response, and thereby reveal scale and shape effects of the covariates, not simply their location shift effect.

Lecture 2. Inference for Quantile Regression

- Inference for the Sample Quantiles
- QR Inference in iid Error Models*
- QR Inference in Heteroscedastic Error Models*
- Classical Rank Tests and the Quantile Regression Dual*

The Precision of Sample Quantiles?

For random samples from a continuous distribution, F, the sample quantiles, $\hat{F}_n^{-1}(\tau)$ are consistent, by the Glivenko-Cantelli theorem. Rates of convergence and precision are governed by the density near the quantile of interest, if it exists.
Note that differentiating the identity: $F(F^{-1}(t)) = t$, yields,

$$\frac{d}{dt}F(F^{-1}(t)) = f(F^{-1}(t))\frac{d}{dt}F^{-1}(t) = 1$$

thus, provided of course that $f(F^{-1}(t)) > 0$,

$$\frac{d}{dt}F^{-1}(t) = \frac{1}{f(F^{-1}(t))}.$$

So, limiting normality of $\hat{F}_n$ and the δ-method imply limiting normality of the sample quantiles with $\sqrt{n}$ rate and variance proportional to $f^{-2}(F^{-1}(t))$.

Inference for the Sample Quantiles

Minimizing $\sum_{i=1}^n \rho_\tau(y_i - \xi)$ consider

$$g_n(\xi) = -n^{-1}\sum_{i=1}^n \psi_\tau(y_i - \xi) = n^{-1}\sum_{i=1}^n (I(y_i < \xi) - \tau).$$

By convexity of the objective function,

$$\{\hat{\xi}_\tau > \xi\} \Leftrightarrow \{g_n(\xi) < 0\}$$

and the DeMoivre-Laplace CLT yields, expanding F,

$$\sqrt{n}(\hat{\xi}_\tau - \xi) \rightsquigarrow \mathcal{N}(0, \omega^2(\tau, F))$$

where $\omega^2(\tau, F) = \tau(1-\tau)/f^2(F^{-1}(\tau))$. Classical Bahadur-Kiefer representation theory provides further refinement of this result.

Some Gory Details

Instead of a fixed $\xi = F^{-1}(\tau)$ consider,

$$\mathbb{P}\{\hat{\xi}_n > \xi + \delta/\sqrt{n}\} = \mathbb{P}\{g_n(\xi + \delta/\sqrt{n}) < 0\}$$

where $g_n \equiv g_n(\xi + \delta/\sqrt{n})$ is a sum of iid terms with

$$\begin{aligned} \mathbb{E}g_n &= \mathbb{E}n^{-1}\sum_{i=1}^{n}(I(y_i < \xi + \delta/\sqrt{n}) - \tau) \\ &= F(\xi + \delta/\sqrt{n}) - \tau \\ &= f(\xi)\delta/\sqrt{n} + o(n^{-1/2}) \\ &\equiv \mu_n\delta + o(n^{-1/2}) \end{aligned}$$

$$\mathbb{V}g_n = \tau(1-\tau)/n + o(n^{-1}) \equiv \sigma_n^2 + o(n^{-1}).$$

Thus, by (a triangular array form of) the DeMoivre-Laplace CLT,

$$\mathbb{P}(\sqrt{n}(\hat{\xi}_n - \xi) > \delta) = \Phi((0 - \mu_n\delta)/\sigma_n) \equiv 1 - \Phi(\omega^{-1}\delta)$$

where $\omega = \mu_n/\sigma_n = \sqrt{\tau(1-\tau)}/f(F^{-1}(\tau))$.

Finite Sample Density for QR

Let $h \in \mathcal{H}$ index the $\binom{n}{p}$ p-element subsets of $\{1, 2, \ldots, n\}$ and $X(h), y(h)$ denote corresponding submatrices and vectors of X and y.

Lemma: $\hat{\beta} = b(h) \equiv X(h)^{-1}y(h)$ is the τth regression quantile iff $\xi_h \in \mathcal{C}$ where $\xi_h = \sum_{i \notin h} \psi_\tau(y_i - x_i\hat{\beta})x_i^\top X(h)^{-1}$, $\mathcal{C} = [\tau - 1, \tau]^p$, and $\psi_\tau(u) = \tau - I(u < 0)$.

Theorem: (Koenker and Bassett, 1978) In the linear model with iid errors, $\{u_i\} \sim F, f$, the density of $\hat{\beta}(\tau)$ is given by

$$\begin{aligned} g(b) = &\sum_{h\in\mathcal{H}} \prod_{i\in h} f(x_i^\top(b - \beta(\tau)) + F^{-1}(\tau)) \\ &\cdot P(\xi_h(b) \in C)|\det(X(h))| \end{aligned}$$

Asymptotic behavior of $\hat{\beta}(\tau)$ follows by (painful) consideration of the limiting form of this density, see also Goh and Knight (2009).

Asymptotic Theory of Quantile Regression I

In the classical linear model,

$$y_i = x_i\beta + u_i$$

with u_i iid from df F, with density $f(u) > 0$ on its support $\{u | 0 < F(u) < 1\}$, the joint distribution of $\sqrt{n}(\hat{\beta}_n(\tau_i) - \beta(\tau_i))_{i=1}^m$ is asymptotically normal with mean 0 and covariance matrix $\Omega \otimes D^{-1}$. Here $\beta(\tau) = \beta + F_u^{-1}(\tau)e_1$, $e_1 = (1, 0, \ldots, 0)^\top$, $x_{1i} \equiv 1$, $n^{-1}\sum x_i x_i^\top \to D$, a positive definite matrix, and

$$\Omega = ((\tau_i \wedge \tau_j - \tau_i\tau_j)/(f(F^{-1}(\tau_i))f(F^{-1}(\tau_j)))_{i,j=1}^m.$$

Asymptotic Theory of Quantile Regression II

When the response is conditionally independent over i, but not identically distributed, the asymptotic covariance matrix of $\zeta(\tau) = \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$ is somewhat more complicated. Let $\xi_i(\tau) = x_i\beta(\tau)$, $f_i(\cdot)$ denote the corresponding conditional density, and define,

$$\begin{aligned} J_n(\tau_1, \tau_2) &= (\tau_1 \wedge \tau_2 - \tau_1\tau_2)n^{-1}\sum_{i=1}^{n} x_i x_i^\top, \\ H_n(\tau) &= n^{-1}\sum x_i x_i^\top f_i(\xi_i(\tau)). \end{aligned}$$

Under mild regularity conditions on the $\{f_i\}$'s and $\{x_i\}$'s, we have joint asymptotic normality for $(\zeta(\tau_i), \ldots, \zeta(\tau_m))$ with covariance matrix

$$V_n = (H_n(\tau_i)^{-1}J_n(\tau_i, \tau_j)H_n(\tau_j)^{-1})_{i,j=1}^m.$$

Making Sandwiches

The crucial ingredient of the QR Sandwich is the quantile density function $f_i(\xi_i(\tau))$, which can be estimated by a difference quotient. Differentiating the identity: $F(Q(t)) = t$ we get

$$s(t) = \frac{dQ(t)}{dt} = \frac{1}{f(Q(t))}$$

sometimes called the "sparsity function" so we can compute

$$\hat{f}_i(x_i^\top \hat{\beta}(\tau)) = 2h_n / (x_i^\top (\hat{\beta}(\tau + h_n) - \hat{\beta}(\tau - h_n))$$

with $h_n = O(n^{-1/3})$. Prudence suggests a modified version:

$$\tilde{f}_i(x_i^\top \hat{\beta}(\tau)) = \max\{0, \hat{f}_i(x_i^\top \hat{\beta}(\tau))\}.$$

Various other strategies can be employed including a variety of bootstrapping options. More on this in Lecture 3.

Testing Equality of "Slopes"

Wald-type inference with QR sandwiches is applicable to a wide variety of hypotheses. Frequently, it is of interest to test whether covariate (treatment) effects differ across quantiles. In the R package `quantreg` such tests and many others can be accomplished with the command `anova`, e.g.

```
> data(engel)
> f1 <- rq(foodexp ~ income, tau = .25, data = engel)
> f2 <- rq(foodexp ~ income, tau = .5, data = engel)
> f3 <- rq(foodexp ~ income, tau = .75, data = engel)
> anova(f1,f2,f3)
Quantile Regression Analysis of Deviance Table

Model: foodexp ~ income
Joint Test of Equality of Slopes: tau in {  0.25 0.5 0.75  }

  Df Resid Df F value    Pr(>F)
  1  2      703  15.557 2.449e-07
  ---
```


Rank Based Inference for Quantile Regression

- Ranks play a fundamental *dual* role in QR inference, via a generalization Hájek rankscore functions,
- Classical rank tests for the p-sample problem can thereby be extended to regression settings,
- Rank tests play the role of Rao (score) tests for quantile regression,
- Rank tests can be inverted to produce confidence intervals (regions).

Duality of Ranks and Quantiles

We have seen that quantiles may be *defined* as

$$\hat{\xi}(\tau) = \text{argmin} \sum \rho_\tau(y_i - \xi),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$. This can be formulated as a linear program whose dual solution

$$\hat{a}(\tau) = \arg\max\{y^\top a | 1_n^\top a = (1-\tau)n, a \in [0,1]^n\}$$

generates the Hájek rankscore functions.

Reference: Gutenbrunner and Jurečková (1992).

Regression Quantiles and Rank Scores

$$\hat{\beta}_n(\tau) = \text{argmin}_{b \in R^p} \sum \rho_\tau(y_i - x_i^\top b)$$

$$\hat{a}_n(\tau) = \arg\max_{a \in [0,1]^n} \{y^\top a | X^\top a = (1-\tau) X^\top 1_n\}$$

$x^\top \hat{\beta}_n(\tau)$ — Estimates $Q_Y(\tau|x)$
Piecewise constant on $[0,1]$.
For $X = 1_n$, $\hat{\beta}_n(\tau) = \hat{F}_n^{-1}(\tau)$.

$\{\hat{a}_i(\tau)\}_{i=1}^n$ — Regression rankscore functions
Piecewise linear on $[0,1]$.
Generalized to regression via the dual formulation.
For $X = 1_n$, $\hat{a}_i(\tau)$ are the Hajek functions.

Regression Rankscore "Residuals"

The Wilcoxon rankscores,

$$\tilde{u}_i = \int_0^1 \hat{a}_i(t) dt$$

play the role of quantile regression residuals. For each observation y_i they answer the question: on which quantile does y_i lie? The $\tilde{u}_i$ satisfy an orthogonality restriction:

$$X^\top \tilde{u} = X^\top \int_0^1 \hat{a}(t) dt = n\bar{x} \int_0^1 (1-t) dt = n\bar{x}/2.$$

This is something like the $X^\top \hat{u} = 0$ condition for OLS. Note that if the X is "centered" then $\bar{x} = (1, 0, \cdots, 0)$. The $\tilde{u}$ vector is approximately uniformly "distributed;" in the one-sample setting $u_i = (R_i + 1/2)/n$ so they are obviously "too uniform."

Regression Rankscore Tests

$$Y = X\beta + Z\gamma + u$$

$$H_0 : \gamma = 0 \text{ versus } H_n : \gamma = \gamma_0/\sqrt{n}$$

Given the regression rank score process for the restricted model,

$$\hat{a}_n(\tau) = \arg\max \left\{ Y^\top a \,|\, X^\top a = (1-\tau) X^\top 1_n \right\}$$

A test of H_0 is based on the linear rank statistics,

$$\hat{b}_n = \int_0^1 \hat{a}_n(t)\, d\varphi(t)$$

Choice of the score function φ permits test of location, scale or (potentially) other effects.

Regression Rankscore Tests

Theorem: (Gutenbrunner et al., 1993) Under H_n and regularity conditions, the test statistic $T_n = S_n^\top Q_n^{-1} S_n$ where
$S_n = (Z - \hat{Z})^\top \hat{b}_n, \quad \hat{Z} = X(X^\top X)^{-1} X^\top Z, \, Q_n = n^{-1}(Z - \hat{Z})^\top Z - \hat{Z})$

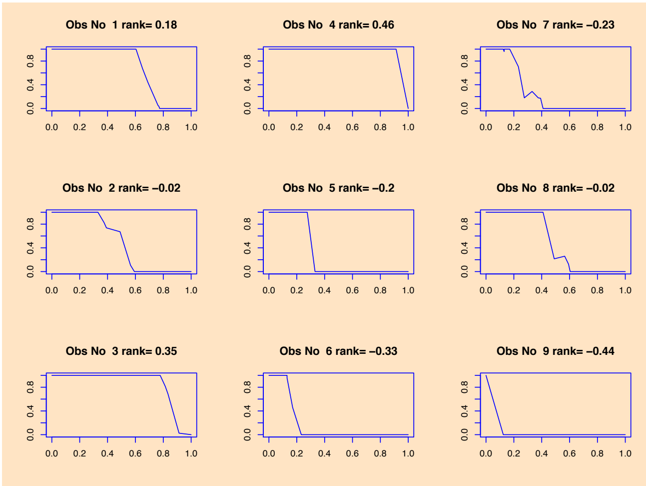
$$T_n \rightsquigarrow \chi_q^2(\eta)$$

where

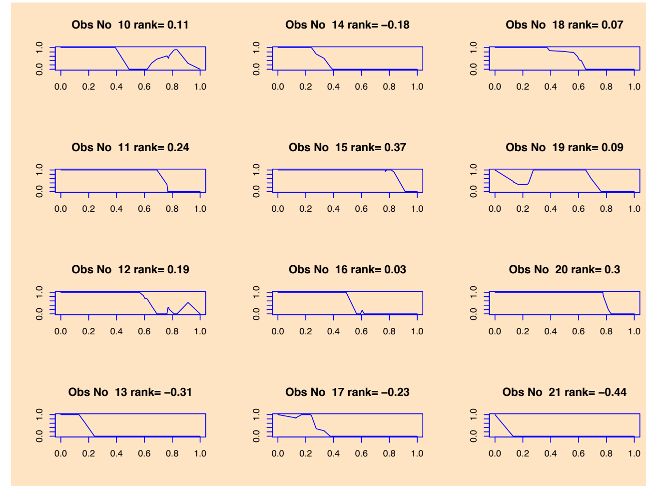
$$\eta^2 = \omega^2(\varphi, F) \gamma_0^\top Q \gamma_0$$

$$\omega(\varphi, F) = \int_0^1 f(F^{-1}(t))\, d\varphi(t)$$

Regression Rankscores for Stackloss Data



Regression Rankscores for Stackloss Data



Inversion of Rank Tests: Confidence Intervals

For the scalar γ case and using the score function

$$\varphi_\tau(t) = \tau - I(t < \tau)$$

$$\hat{b}_{ni} = -\int_0^1 \varphi_\tau(t) d\hat{a}_{ni}(t) = \hat{a}_{ni}(\tau) - (1-\tau)$$

with $\bar{\varphi} = \int_0^1 \varphi_\tau(t)dt = 0$ and $A^2(\varphi_\tau) = \int_0^1 (\varphi_\tau(t) - \bar{\varphi})^2 dt = \tau(1-\tau)$. Thus, a test of the hypothesis $H_0 : \gamma = \xi$ may be based on $\hat{a}_n$ from solving,

$$\max\{(y - x_2\xi)^\top a | X_1^\top a = (1-\tau)X_1^\top 1, a \in [0,1]^n\}$$

and the fact that

$$S_n(\xi) = n^{-1/2} x_2^\top \hat{b}_n(\xi) \rightsquigarrow \mathcal{N}(0, A^2(\varphi_\tau) q_n^2).$$

Inversion of Rank Tests: Confidence Intervals

That is, we may compute

$$T_n(\xi) = S_n(\xi)/(A(\varphi_\tau) q_n),$$

where $q_n^2 = n^{-1} x_2^\top (I - X_1(X_1^\top X_1)^{-1} X_1^\top) x_2$, and reject H_0 if $|T_n(\xi)| > \Phi^{-1}(1 - \alpha/2)$.

Inverting this test, that is finding the interval of ξ's such that the test fails to reject yields a confidence interval for the parameter γ. Unlike the Wald type inference it delivers asymmetric intervals. This is the default approach to parametric inference in R package `quantreg` for problems of modest sample size.

Bootstrap Inference for QR

There are several bootstrapping methods for QR inference

- Standard pairwise Xy
- Gradient resampling a la Parzen et al. (1994)
- Markov chain marginal bootstrap of Kocherginsky et al. (2005)
- Weighted bootstrap of Bose and S. (2003)
- Wild bootstrap of Feng et al. (2011)
- Wild bootstrap for clustered data Hagemann (2014)

Four Concluding Comments about Inference

- Asymptotic inference for quantile regression poses some statistical challenges since it involves elements of nonparametric density estimation, but this shouldn't be viewed as a major obstacle.
- Classical rank statistics and Hájek 's rankscore process are closely linked via Gutenbrunner and Jurečková 's regression rankscore process, providing an attractive approach to many inference problems while avoiding density estimation.
- Inference on the quantile regression process can be conducted with the aid of Khmaladze's extension of the Doob-Meyer construction.
- Resampling offers many further lines of development for inference in the quantile regression setting.

Lecture 3. Computation and Examples

- Estimation in R and Examples
- Inference in R and Examples
- Estimation and Inference in SAS
- Nonparametric Quantile Regression in R

Estimation in Linear Quantile Regression

Linear quantile regression model:

$$y_i = \mathbf{x}_i^T \beta(\tau) + e_i, \quad Q_{e_i|\mathbf{x}_i}(\tau) = 0.$$

Estimator:

$$\hat{\beta}(\tau) = \text{argmin}_\beta \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta).$$

Software

- R package `quantreg`
- SAS/STAT PROC QUANTREG
- STATA: `www.stata.com`
- SPLUS
- Mathematica: `http://mathematicaforprediction.wordpress.com`
- Matlab code for quantile regression:
  `www.econ.uiuc.edu/~roger/research/rq/rq.m`
  `www.stat.psu.edu/~dhunter/code/qrmatlab`

Estimation in R

- R package **quantreg**.
- Install the quantreg package in R
  - From CRAN directly: `install.packages("quantreg")`
  - From local source:

```
install.packages("dir/filename", repos = NULL,
type="source")

e.g. Windows:
install.packages("C:/quantreg_5.05.zip",
repos = NULL, type="source")

Mac:
install.packages("/Users/Documents/
quantreg_5-1.05.tar", repos = NULL,
type="source")
```


Estimation in R

- The syntax of rq() function

```
library(quantreg)
rq(formula, tau=0.5, data, method="br", ...)
```

- formula: model statement, e.g.

$$y \sim x1 + x2$$

- tau: quantile level(s) of interest
  - If tau is a single number $\in (0, 1)$, returns the estimates from regression at this specified quantile level.
  - If tau is a vector, e.g. tau = c(0.25, 0.5, 0.75), returns the estimates from regression at multiple quantiles.
  - If tau $\notin (0, 1)$, returns the entire quantile process.

Estimation in R

- rq methods
  - method="br": Simplex method (default). Efficient for problems with modest size (several thousands). Slow relative to LSE for large problems.
  - method="fn": the Frisch-Newton interior point method. More efficient than simplex for larger sample size.
  - method="pfn": the Frisch-Newton approach with preprocessing. Recommended for very large problems (e.g. $n > 10^5$).
  - method="lasso": the Frisch-Newton approach with lasso penalty. Recommended for very large problems (e.g. $n > 10^5$). See also `rqss`.

Example 1: Engel's Food Expenditure

Linear quantile regression model:

$$Q_\tau(\text{foodexp}|\text{income}) = \alpha_\tau + \beta_\tau \times \text{income}.$$

R code:

```
library(quantreg)
#plot the data
plot(engel)

#estimation at median
f1 = rq(foodexp~income, tau=0.5, data=engel)

#superimpose median regression line
abline(f1,col="red")
```


Example 1: Engel's Food Expenditure

```
#view the estimated coefficients
f1
coef(f1) #or f1$coef

#fit at other quantile levels and plot
taus = c(0.1, 0.25, 0.75, 0.9)
f2 = rq(foodexp~income, taus, data=engel)
f2
for(i in 1:length(taus)){
abline(coef(f2)[,i],col="blue")
}
```


Example 1: Engel's Food Expenditure

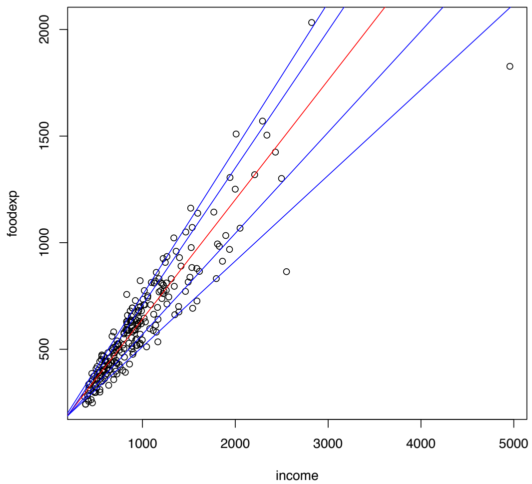
```
> f1
Call:
rq(formula = foodexp ~ income, tau = 0.5, data = engel)

Coefficients:
(Intercept)      income
 81.4822474   0.5601806

> f2
Call:
rq(formula = foodexp ~ income, tau = taus, data = engel)

Coefficients:
              tau= 0.10  tau= 0.25  tau= 0.75  tau= 0.90
(Intercept) 110.1415742 95.4835396 62.3965855 67.3508721
income        0.4017658  0.4741032  0.6440141  0.6862995
```


Example 1: Engel's Food Expenditure



Example 1: Engel's Food Expenditure

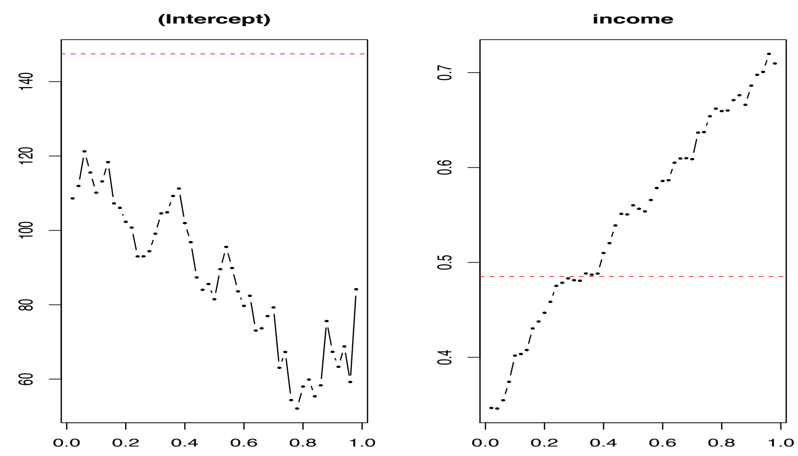
Automatic plots of the estimated quantile coefficients across quantile levels can be obtained by using the R function `plot.rqs`:

```
taus = 1:49/50
fm = rq(foodexp~income, taus, data=engel)
plot(fm, mfrow=c(1,2))

#or plot only the second coefficient (slope)
plot(fm, parm = 2, xlab = "tau", cex = 1, pch = 19)
```

Try the R function `example(rq)` for illustration of quantile regression analysis on the data **stackloss** and **engel**.

Example 1: Engel's Food Expenditure



Compare the Computing Time of Estimation Methods via Simulation

```
library(quantreg)
#a simulated data with one predictor
set.seed(1278991)
n = 1000
x = runif(n)
y = 1 + x + (1+x)*rnorm(n)
system.time(lmfit <- lm(y~x))
system.time(f1 <- rq(y~x, tau=0.5, method="br"))
system.time(f2 <- rq(y~x, tau=0.5, method="fn"))
system.time(f3 <- rq(y~x, tau=0.5, method="pfn"))

> system.time(lmfit <- lm(y~x))
   user  system elapsed
  0.349   0.008   0.445
> system.time(f1 <- rq(y~x, tau=0.5, method="br"))
   user  system elapsed
  0.003   0.001   0.009
> system.time(f2 <- rq(y~x, tau=0.5, method="fn"))
   user  system elapsed
  0.328   0.004   0.352
> system.time(f3 <- rq(y~x, tau=0.5, method="pfn"))
   user  system elapsed
  0.008   0.002   0.014
```


A Simulated Data with $n = 10,000$

```
> set.seed(1278991)
> n = 10000
> x = runif(n)
> y = 1 + x + (1+x)*rnorm(n)
> system.time(lmfit <- lm(y~x))
   user  system elapsed
  0.301   0.016   0.491
> system.time(f1 <- rq(y~x, tau=0.5, method="br"))
   user  system elapsed
  0.304   0.008   0.342
> system.time(f2 <- rq(y~x, tau=0.5, method="fn"))
   user  system elapsed
  0.035   0.002   0.043
> system.time(f3 <- rq(y~x, tau=0.5, method="pfn"))
   user  system elapsed
  0.217   0.009   0.280
```

A Simulated Data with $n = 200,000$

```
> set.seed(1278991)
> n = 200000
> x = runif(n)
> y = 1 + x + (1+x)*rnorm(n)
> system.time(lmfit <- lm(y~x))
   user  system elapsed
  0.955   0.036   1.119
> system.time(f1 <- rq(y~x, tau=0.5, method="br"))
   user  system elapsed
  9.438   0.074   9.516
> system.time(f2 <- rq(y~x, tau=0.5, method="fn"))
   user  system elapsed
  1.031   0.066   1.099
> system.time(f3 <- rq(y~x, tau=0.5, method="pfn"))
   user  system elapsed
  0.562   0.055   0.635
```


A Simulated Data with Sparse Design Matrix

```
set.seed(1278991)
n = 1000; p=500
X = matrix(rbinom(n*p,1,0.5),ncol=p)
beta = c(rep(2,5), rep(0,p-5))
y = 1 + X%*%beta + (1+X[,1])*rnorm(n)
system.time(f1 <- rq(y~X, tau=0.5, method="br"))
system.time(f2 <- rqss(y~X, tau=0.5, method="sfn"))
system.time(f3 <- rq(y~X, tau=0.5, method="lasso",
lambda=10))
plot(f3$coef)

> system.time(f1 <- rq(y~X, tau=0.5, method="br"))
   user  system elapsed
  9.439   0.056   9.494
> system.time(f2 <- rqss(y~X, tau=0.5, method="sfn"))
   user  system elapsed
  4.082   0.202   4.305
> system.time(f3 <- rq(y~X, tau=0.5, method="lasso",
+ lambda=10))
   user  system elapsed
  2.046   0.042   2.110
```


Inference in R

function:

```
rq.object = rq(y~x, tau)
summary.rq(rq.object, se="nid",...)
```

- se="rank": default method, constructs confidence intervals by the inversion of **rank score test** (Gutenbrunner et al., 1993; Koenker, 1994). The default confidence level is 90%, use “alpha=0.05" to obtain 95% confidence intervals.
- se=“nid": based on the asymptotic normality and **direct estimation** of the asymptotic variance assuming non i.i.d. errors.
- se=“iid": similar as “nid" but assumes i.i.d. errors.
- se="ker": based on direct estimation of the asymptotic variance and uses a kernel estimate of the sandwich.

Inference in R

- se="boot": uses **bootstrap** to estimate standard errors.
  - bsmethod=“xy": xy-pair bootstrap.
  - bsmethod=“pwy": resampling the estimating equations (Parzen et al., 1994).
  - bsmethod=“mcmb": Markov chain marginal bootstrap (He and Hu, 2002), relieves the computational burden by reducing a high-dimensional problem to several one-dimensional problems.
  - bmethod = "wxy": uses the generalized bootstrap with unit exponential weights (Bose and S., 2003).
  - bsmethod=“wild": wild bootstrap (Feng et al., 2011).

Engel's—Confidence Band (rank)

- Manually construct 95% pointwise confidence band of the estimated coefficients using the rank score method.

```
taus = 1:49/50
coef = lb = ub = NULL
for(i in 1:length(taus)){
  obj = rq(foodexp~income, tau=taus[i], data=engel)
  f1 = summary(obj, se="rank", alpha=0.05)
  coef = rbind(coef, f1$coef[,1])
  lb = rbind(lb, f1$coef[,2])
  ub = rbind(ub, f1$coef[,3])
}
#intercept
plot(coef[,1]~taus,ylim=range(c(coef[,1],lb[,1],ub[,1])),
ylab="intercept",type="b")
lines(lb[,1]~taus,col="blue",lty=5)
lines(ub[,1]~taus,col="blue",lty=5)

#slope
plot(coef[,2]~taus,ylim=range(c(coef[,2],lb[,2],ub[,2])),
ylab="income",type="b")
lines(lb[,2]~taus,col="blue",lty=5)
lines(ub[,2]~taus,col="blue",lty=5)
```

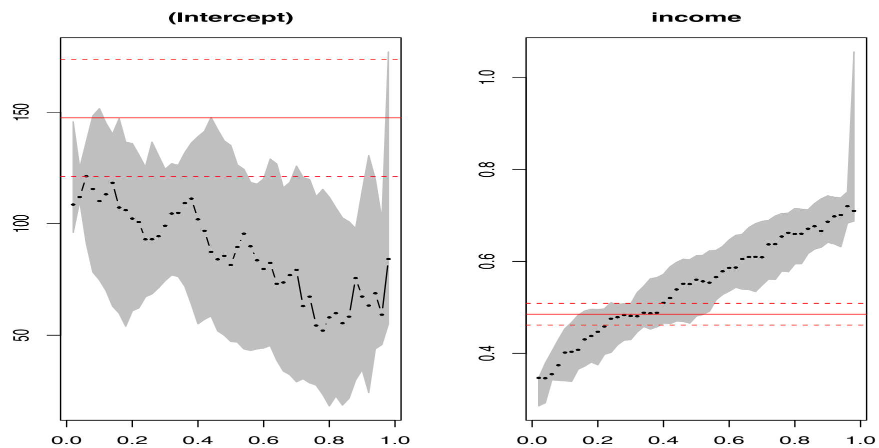

Engel's—Confidence Band (rank)

- Construct 95% pointwise confidence bands automatically by using the R function `plot.summary.rqs()`.

```
taus = 1:49/50
fm = rq(foodexp~income, taus, data=engel)
sfm = summary(fm, se="rank", alpha=0.05)
plot(sfm, mfrow=c(1,2))

#plot the CB for only the slope
plot(sfm, parm = 2, xlab = "tau", cex = 1, pch = 19)
```


Engel's—Confidence Band (rank)

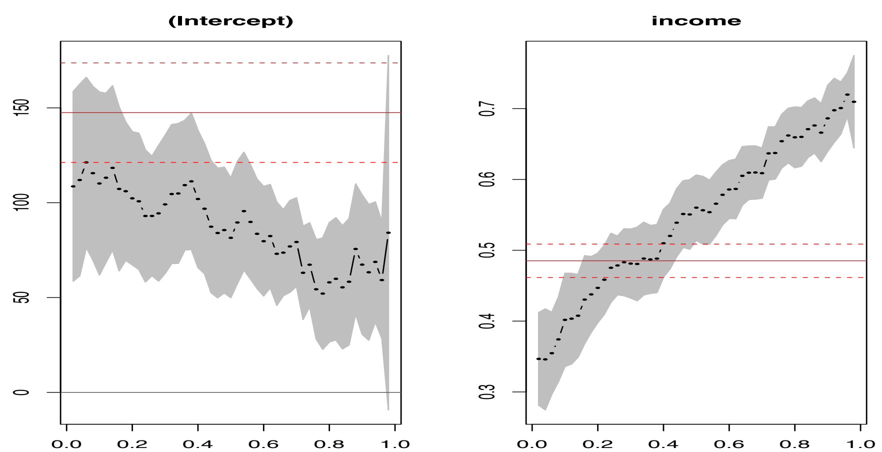


Engel's—Confidence Band (nid)

- Construct 95% pointwise confidence band of the estimated coefficients using the "nid" method.

```
sfm = summary(fm, se="nid")
plot(sfm, mfrow=c(1,2))
```

Engel's—Confidence Band (nid)



Example 2: Birthweight

Data: a random subset of NATALITY1997, 10,000 observations

Variable	Description
weight	Infant's birth weight (grams)
black	Indicator of black mother
married	Indicator of married mother
boy	Indicator of boy
visit	Prenatal visit: 0 = no visit, 1 = visit in 1st, 2 = visit in 2nd, 3 = visit in 3rd trimester
ed	Mother's edu.: 0 = high school, 1 = some college, 2 = college, 3 = less than high school
smoke	Indicator of smoking mother
cigsper	Number of cigarettes smoked per day
momage	Mother's age
mwtgain	Mother's weight gain during pregnancy

Example 2: Infant Birthweight

Quantile regression model:

$$\begin{aligned} Q_\tau(\text{Birthweight}) &= \text{Boy} + \text{Married} + \text{Black} + \text{Mother's Age} \\ &+ \text{Mother's Age}^2 + \text{High School} + \text{Some College} + \text{College} \\ &+ \text{No Prenatal} + \text{Prenatal Second} + \text{Prenatal Third} \\ &+ \text{Smoker} + \text{Cigarette's/Day} + \text{Mother's Weight Gain} \\ &+ \text{Mother's Weight Gain}^2 \end{aligned}$$

Quantile levels considered:

$$\tau = 0.05, 0.1, 0.15, \dots, 0.9, 0.95$$

Estimation—R Code

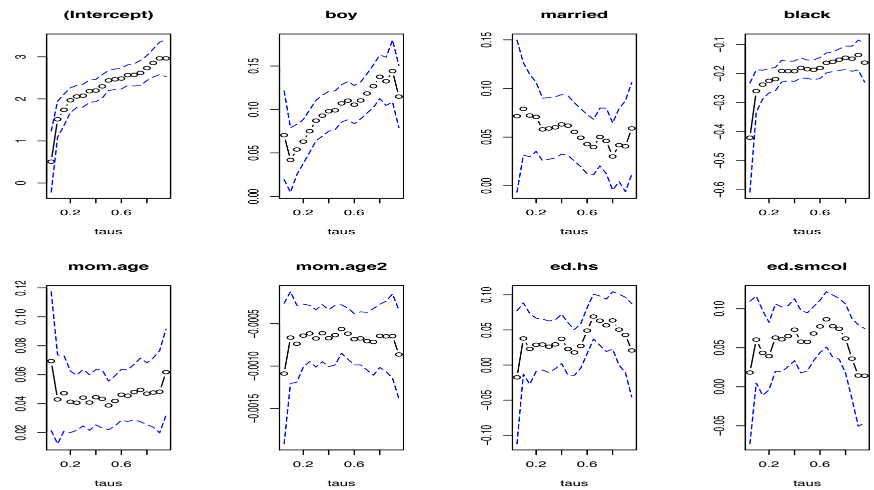
```
bweight = read.csv(
"http://www4.stat.ncsu.edu/~wang/RQ/bweight.csv")
taus=seq(0.05, 0.95, 0.05)
coef = se = NULL
for(i in 1:length(taus)){
obj = rq(weight~ boy + married + black + mom.age +
     mom.age2 + ed.hs + ed.smcol + ed.col + novisit +
     tri2 + tri3 + smoke + cigsper + m.wtgain +
     m.wtgain2, tau=taus[i], data=bweight)
f1 = summary(obj, se="nid")
coef = rbind(coef, f1$coef[,1])
se = rbind(se, f1$coef[,2])
}
lb = coef - 1.96*se
ub = coef + 1.96*se
}
```

Estimation—R Code

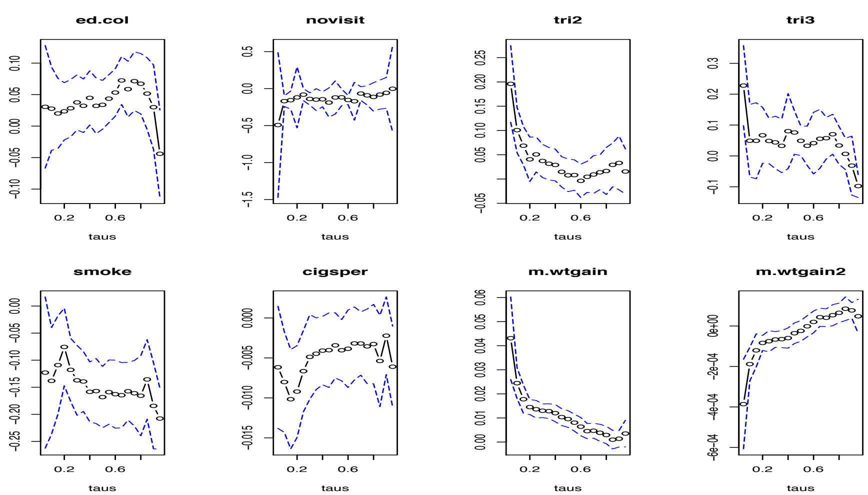
```
par(mfrow=c(2,4))
for(j in 1:8){
plot(coef[,j]~taus,ylim=range(c(coef[,j],lb[,j],ub[,j
main=names(coef(obj))[j],type="b")
lines(lb[,j]~taus,col="blue",lty=5)
lines(ub[,j]~taus,col="blue",lty=5)
}

for(j in 9:16){
plot(coef[,j]~taus,ylim=range(c(coef[,j],lb[,j],ub[,j
main=names(coef(obj))[j],type="b")
lines(lb[,j]~taus,col="blue",lty=5)
lines(ub[,j]~taus,col="blue",lty=5)
}
```


Confidence Band (nid)



Confidence Band (nid)



Estimation in SAS

- Package: SAS/STAT PROC QUANTREG
- **Basic syntax**
  PROC QUANTREG DATA = sas-data-set <options> **CI** = <NONE|RANK|...> **ALPHA**=value;
  BY variables;
  CLASS variables;
  MODEL response = independents </options>;
  RUN;

To specify the quantile level

- Use the option QUANTILE in the MODEL statement
  MODEL Y = X / QUANTILE = <number list | ALL>;

Choice of quantile(s)	Example
A single quantile	QUANTILE = 0.25
Multiple quantiles	QUANTILE = 0.25 0.5 0.75
Entire Quantile process	QUANTILE = ALL

- Default value is 0.5, corresponding to the median.

To specify the algorithm

- Use the option **ALGORITHM** in the **PROC QUANTREG** statement

Method	Option
Simplex	ALGORITHM = SIMPLEX
Interior point	ALGORITHM = INTERIOR
Interior point with preprocessing	ALGORITHM = INTERIOR PP
Smoothing	ALGORITHM = SMOOTHING

- The default is Simplex algorithm.

Default Output

- **Model information**: report the name of the data set and the response variable, the number of covariates, the number of observations, algorithm of optimization and the method for confidence intervals.
- **Summary statistics**: report the sample mean and standard deviation, sample median, MAD and interquartile range for each variable included in the MODEL statement.
- **Quantile objective function**: report the quantile level to be estimated and the optimized objective function.
- **Parameter Estimates**: report the estimated coefficients and their 95% confidence intervals.

Nonparametric Quantile Regression in R

- Locally polynomial method

```
lprq(y~x, h= , tau= )
```

- B-spline method

```
rq(y~bs(x, knots= ), tau= )
```

- Smoothing spline (penalty) method

```
rqss(y~qss(x, constraint="N"),  tau= )

#options for constraint:
"N","I","D","V","C" "VI","VD","CI","CD"
```

Example 3: Growth Chart

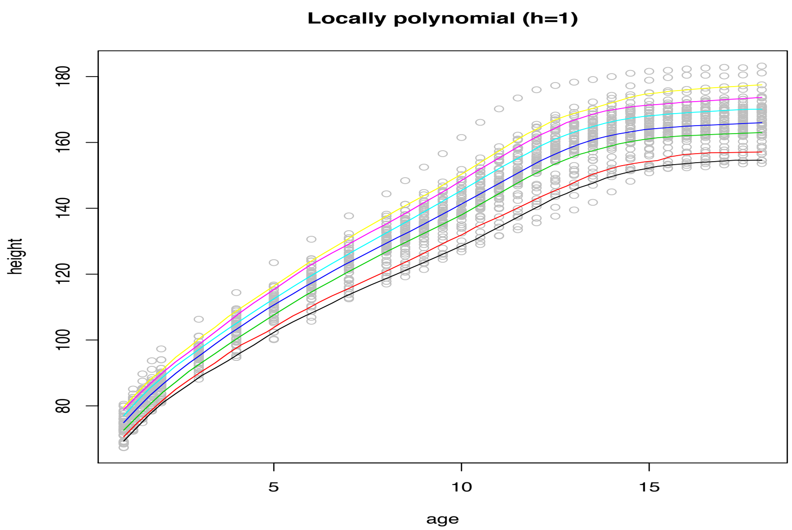
- Data: Berkeley growth data (Tuddenham and Snyder, 1954)
- Number of female subjects: 54
- Measurements: longitudinal height measurements collected at 31 time points from age 0 to age 18, specifically
  - quarterly between ages 0 and 2;
  - yearly between ages 2 and 8;
  - semi-yearly between ages 8 and 21.

Growth Chart—R code

```
growth = read.csv(
"http://www4.stat.ncsu.edu/~wang/RQ/growth.csv")
attach(growth)
taus = c(0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95)

#locally polynomial (kernel) method
plot(height~age,col="grey", main="Locally polynomial")
for(i in 1:length(taus)){
fit = lprq(age, height, h=1, taus[i])
lines(fit$fv~fit$xx, col=i)
}
```


Growth Chart (lprq)

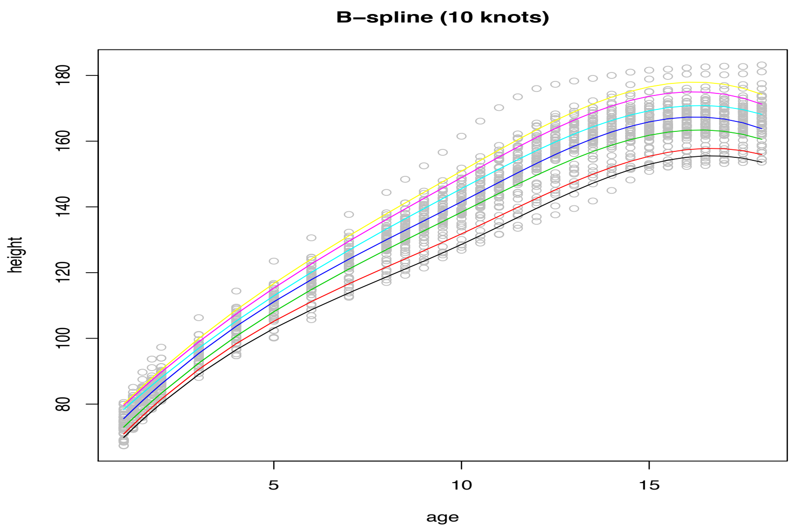


Growth Chart—R code

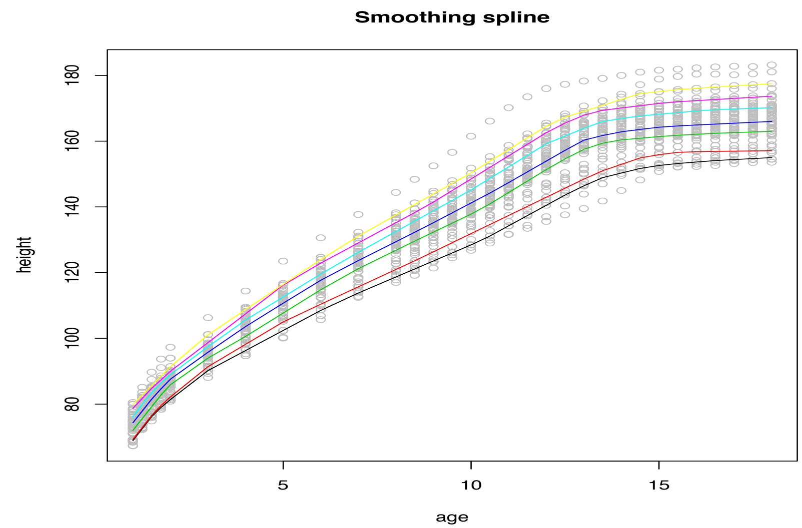
```
#B-spline method
plot(height~age,col="grey", main="B-spline (10 knots)")
for(i in 1:length(taus)){
fit = rq(height~bs(age,knots=10), tau=taus[i])
lines(fit$fitted[1:31]~age[1:31], col=i)
}


#smoothing spline method
plot(height~age,col="grey", main="Smoothing spline")
for(i in 1:length(taus)){
fit = rqss(height~qss(age,constraint="N"), tau=taus[i])
fit = c(fit$coef[1], fit$coef[1]+fit$coef[-1])
lines(fit~age[1:31], col=i)
}
```

Growth Chart (B-spline)



Growth Chart (rqss)



Lecture 4. Censored QR and Survival Analysis

- Beyond the Transformation Model
- Quantile Regression under Censorship
  - Fixed Censoring
  - Random Censoring
- Some One-sample Asymptotics
- Conclusions

Transformation Model

A wide variety of survival analysis models (e.g. Cox, proportional-odds), following Doksum and Gasko (1990), may be written as,

$$h(T_i) = x_i^\top \beta + u_i,$$

- T_i: an observed survival time;
- h: a monotone transformation;
- x_i: a vector of covariates;
- β: an unknown parameter vector;
- $\{u_i\}$: $i.i.d. \sim F$.

Accelerated failure time model:

$$\log(T_i) = x_i^\top \beta + u_i$$

Beyond the Transformation Model

The common feature of all these models is that after transformation of the observed survival times we have:

- a pure **location-shift**, iid-error regression model;
- covariate effects shift the center of the distribution of $h(T)$;
- but covariates **cannot affect scale, or shape** of this distribution.

Quantile Regression Model

Quantile regression model:

$$Q_{h(T_i)}(\tau|\mathbf{x}_i) = \mathbf{x}_i^T \beta(\tau),$$

where $h(\cdot)$ is a monotone transformation. By the equivariance property of quantile regression to monotone transformation (suppose h is increasing),

$$Q_{T_i}(\tau|\mathbf{x}_i) = h^{-1}\left\{\mathbf{x}_i^T \beta(\tau)\right\}.$$

Claim I:

- quantile regression allows the covariate to affect **not only location but also scale and shape** of the conditional distribution;
- interpretation is simpler.

Quantile Regression under Censorship

WLLG assume T_i is the transformed survival/duration.

Data: $(\mathbf{x}_i, Y_i, \delta_i)$, $i = 1, \cdots, n$, where

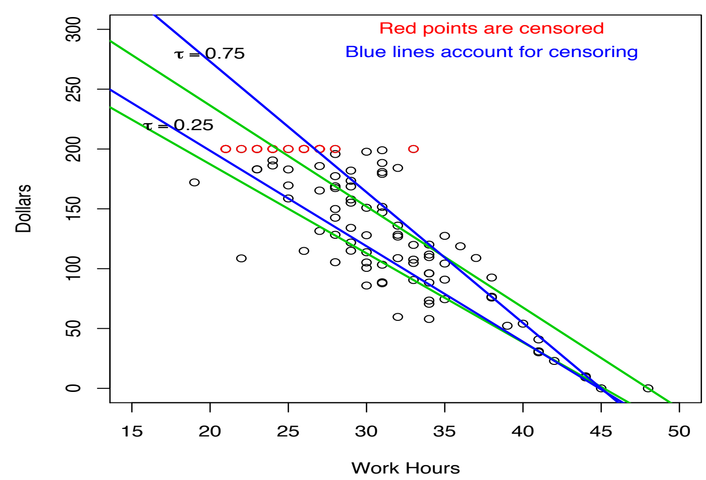
$$Y_i = \min(T_i, C_i), \quad \delta_i = I(T_i \leqslant C_i).$$

Censored quantile regression:

$$Q_{T_i}(\tau|\mathbf{x}_i) = \mathbf{x}_i^T \beta_0(\tau).$$

Why Censored Quantile Regression?

Example: Average Weekly Earnings v.s. Study Hours



Identifiability under Censoring

- Conditional mean $E(T|X)$ is **not identifiable**.
- But the conditional quantiles $Q_T(\tau|\mathbf{x})$ are **identifiable** for some τ.

40% right censoring (red) at 200.
Identifiable quantile region: $\tau \in (0, 0.6)$.

Claim II: quantile regression has **better identifiability under censorship.**

Censored Quantile Regression

- **Fixed censoring**: the censoring times C_i are observed or known for all observations, even for those subjects that are not censored. WLOG assume $C_i = C$. Examples:
  - viral load of HIV patients;
  - antibody concentration in blood;
  - age or salary in survey studies.
- **Random censoring**: censoring points are unknown for uncensored observations, more common in biomedical studies.

Powell's Approach for Fixed Censoring

- **Rationale**: quantiles are equivariant to monotone transformations so

$$Q_T(\tau|\mathbf{x}_i) = \mathbf{x}_i^T \beta_0(\tau), \quad Y_i = \min(T_i, C)$$
$$\Rightarrow \quad Q_Y\{\tau|\mathbf{x}_i\} = \min\{\mathbf{x}_i^T \beta_0(\tau), C\}.$$

- Powell's estimator:

$$\hat{\beta}(\tau) = \text{argmin}_\beta \sum_{i=1}^{n} \rho_\tau\{Y_i - \min(\mathbf{x}_i^T \beta, C)\}.$$

- Powell's approach estimates truncated conditional quantile functions (nonlinear in parameters).

References: Powell (1984, 1986); Fitzenberger (1997); Koenker and Park (1996).

About Powell's Estimator

- Semiparametric optimality for Powell's estimator was claimed in Newey and Powell (1990)
  - but the optimality result is far from optimal.
- **Computational challenges** (Fitzenberger, 1997; Zhou, 2006; Stengos and Wang, 2007)
  - non-convex objective function;
  - easy to get stuck at a local minimum;
  - exponential growth in computation

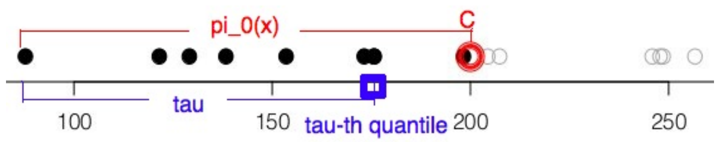
Informative Subset Estimation (ISUB)

Tang et al. (2012)

- Powell's estimator is asymptotically equivalent to

$$\text{argmin}_{\beta} \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{x}_i^T \beta) I\{\mathbf{x}_i^T \beta_0(\tau) < C\}. \quad (1)$$

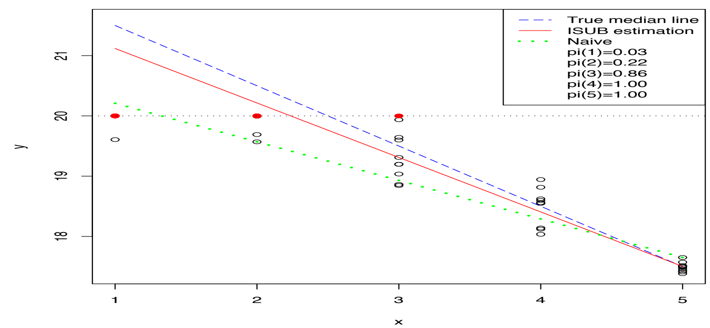
- **Idea:** apply standard QR on $(\mathbf{x}_i, Y_i)$ to the informative subset: $\{i : \mathbf{x}_i^T \beta_0(\tau) < C\}$.
- $Q_\tau(T|\mathbf{x}) = \mathbf{x}^T \beta_0(\tau) < C \Leftrightarrow \pi_0(\mathbf{x}) = P(\delta = 1|\mathbf{x}) > \tau$



ISUB

Idea: apply standard QR on (x_i, Y_i) to the informative subset:

$$\{i : \mathbf{x}_i^T \beta_0(\tau) < C\} = \{i : \pi_0(\mathbf{x}_i) > \tau\}.$$



$\pi_0(x_i) = P(\delta_i = 0|\mathbf{x}_i)$ can be estimated by parametric or nonparametric regression on $(1 - \delta_i, \mathbf{x}_i)$, e.g. local logistic regression, generalized additive model.

Random Censoring

Random censoring: censoring points are unobserved for uncensored observations.

Two Different Assumptions

- **Assumption A**: C is independent of T and $\mathbf{X}$.
- **Assumption B**: C and T are independent conditional on $\mathbf{X}$.

Common Approach under Assumption A

Idea: re-weight the quantile estimating equation accounting for censoring.

- Ying et al. (1995)

$$\sum_{i=1}^{n} \mathbf{x}_i \left\{ \frac{I(Y_i > \mathbf{x}_i^T \beta)}{\hat{G}(\mathbf{x}_i^T \beta)} - (1-\tau) \right\} \approx 0, \quad (2)$$

where G is the Kaplan-Meier estimate of the survival function of C_i. **Rationale:**

$$\begin{aligned} & P\{Y_i > \mathbf{x}_i^T \beta_0(\tau)|\mathbf{x}_i\} = P\{\min(T_i, C_i) > \mathbf{x}_i^T \beta_0(\tau)|\mathbf{x}_i\} \\ = \ & P\{T_i > \mathbf{x}_i^T \beta_0(\tau)|\mathbf{x}_i\} P\{C_i > \mathbf{x}_i^T \beta_0(\tau)|\mathbf{x}_i\} \\ = \ & (1-\tau) G\{\mathbf{x}_i^T \beta_0(\tau)\}, \end{aligned}$$

thus the estimating function in (2) is unbiased if G is known.

Common Approach under Assumption A

- Bang and Tsiatis (2002):

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(Y_i)} \mathbf{x}_i \{ I(Y_i < \mathbf{x}_i^T \beta) - \tau \} \approx 0. \tag{3}$$

Rationale:

$$\begin{aligned} & E\left[\frac{\delta_i}{G(Y_i)} \{ I(Y_i < \mathbf{x}_i^T \beta) - \tau \} | \mathbf{x}_i \right] \\ = \; & E\left(E\left[\frac{I(T_i < C_i)}{G(T_i)} \{ I(T_i < \mathbf{x}_i^T \beta) - \tau \} | \mathbf{x}_i, T_i \right] \right) \\ = \; & 0 \end{aligned}$$

when $\beta = \beta_0(\tau)$.

Common Approach under Assumption B

Idea: distribute mass of censored observations to the right

$$\min \sum_{i=1}^{n} \left[w_i \rho_\tau (Y_i - \mathbf{x}_i^T \beta) + (1 - w_i) \rho_\tau (+\infty - \mathbf{x}_i^T \beta) \right],$$

where each right censored observation is split into two points

- at Y_i with mass point w_i
- at infinity with mass $1 - w_i$

Reference: Portnoy (2003); Wang and Wang (2009).

Redistribution-of-Mass: How to Distribute?

When there is **no censoring**, $Y_i = T_i$ and $\beta_0(\tau)$ can be estimated by minimizing

$$S_n(\beta) = n^{-1} \sum_{i=1}^{n} \rho_\tau (T_i - \mathbf{x}_i^T \beta), \tag{4}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$. The minimizer of $S_n(\beta)$ is a root of the estimating equation

$$D_n(\beta) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \left\{ \tau - I(T_i - \mathbf{x}_i^T \beta \leqslant 0) \right\} \approx 0. \tag{5}$$

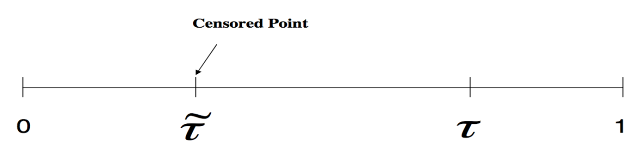
Here $D_n(\beta)$ is the gradient function.

Redistribution-of-Mass: How to Distribute?

Note: the gradient depends only on the signs of $T_i - \mathbf{x}_i^T \beta_0(\tau)$.

- Uncensored: $w_i = 1$.
- **Censored and not yet crossed** (above the τth quantile): i.e., $Y_i = C_i > \mathbf{x}_i^T \beta_0(\tau)$. Treat it as uncensored: $w_i = 1$.
- **Censored and crossed**: $\delta_i = 0$ and $Y_i = C_i < \mathbf{x}_i^T \beta_0(\tau)$, i.e. $\tilde{\tau}_i \doteq F(C_i | \mathbf{x}_i) < \tau$,

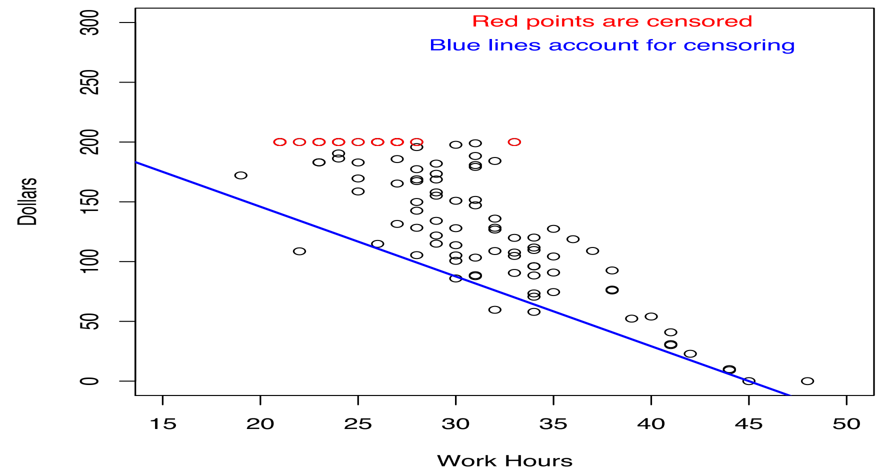
$$E\left\{ I(T_i - \mathbf{x}_i^T \beta_0(\tau) < 0) | T_i > C_i, C_i, \mathbf{x}_i \right\} = \frac{\tau - \tilde{\tau}_i}{1 - \tilde{\tau}_i} \doteq w_i.$$



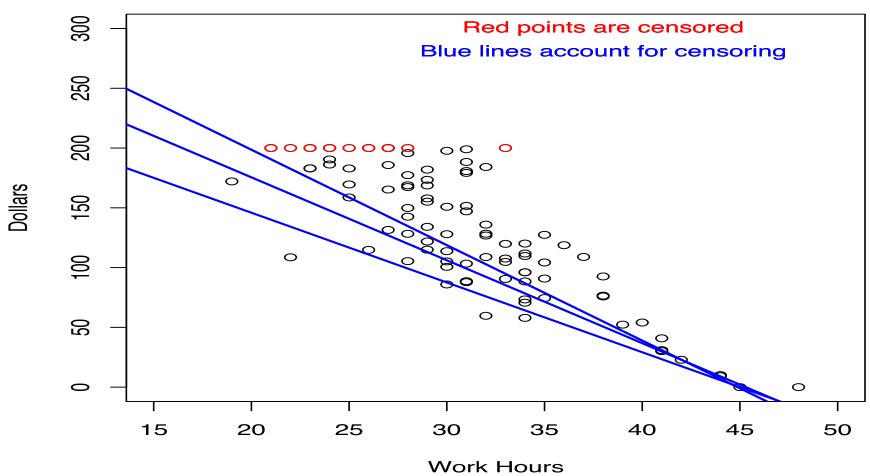
Portnoy's (2003) Approach

- How to estimate $\tilde{\tau}_i = P(T_i \leqslant C_i | \mathbf{x}_i)$: the quantile level at which the conditional quantile crosses C_i?
- Portnoy's approach: estimate quantiles at a fine grid starting from $\tau = 0$ and then move up step by step.

Demonstration: start at $\tau = 0.05$



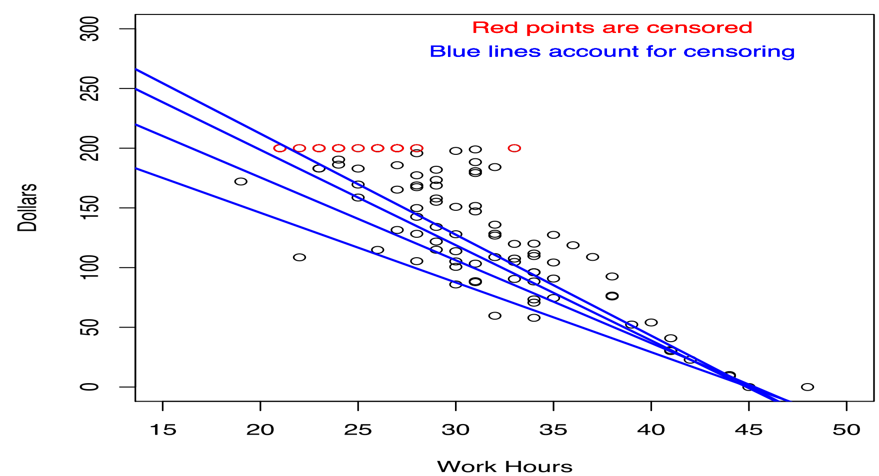
Demonstration: $\tau = 0.05, 0.15, 0.25$



Demonstration: $\tau = 0.05, 0.15, 0.25, 0.35$

For the left two censored and crossed points:

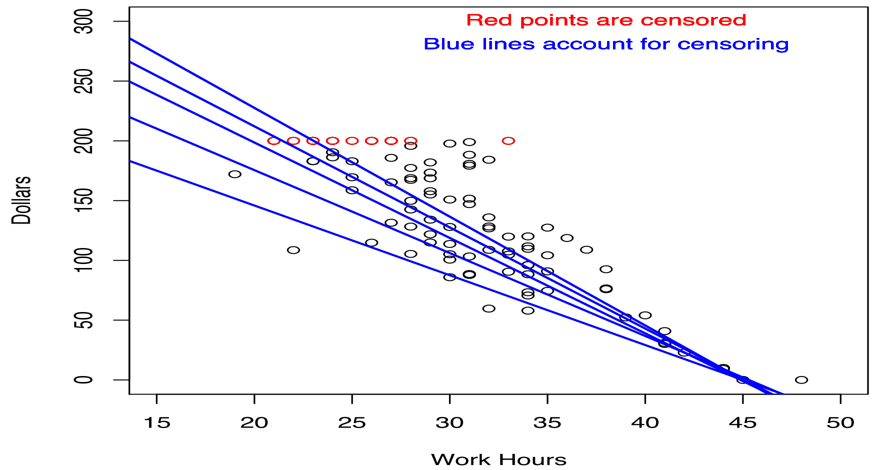
$$\tilde{\tau}_i = 0.35, \quad w_i = \frac{\tau - 0.35}{1 - 0.35} \text{ for } \tau > 0.35.$$



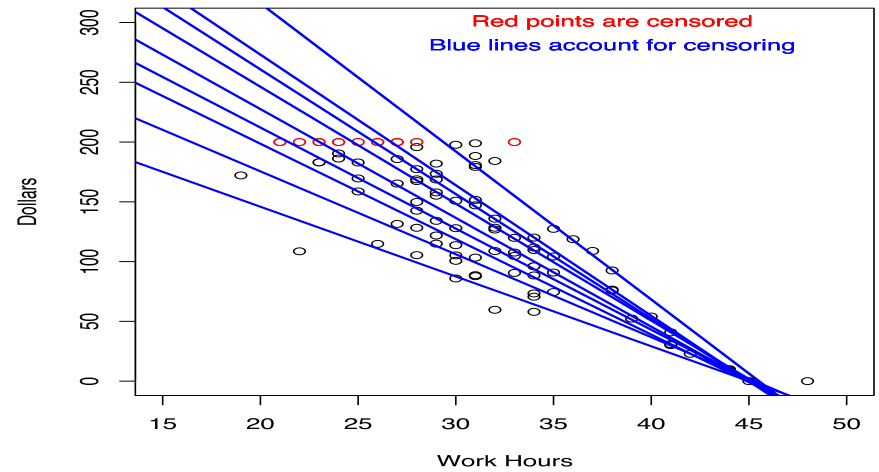
Demonstration: $\tau = 0.05, 0.15, 0.25, 0.35, 0.45$

For the third censored and crossed point:

$$\tilde{\tau}_i = 0.45, \quad w_i = \frac{\tau - 0.45}{1 - 0.45} \text{ for } \tau > 0.45.$$



Demonstration: $\tau = \cdots, 0.75, 0.85$



About Portnoy's Approach

- Efron's (1967) redistribution-of-mass idea: all observations are assigned weights depending on whether they are **uncensored, censored but not yet crossed**, or **censored and crossed**.
- Reduces to Kaplan Meier's estimator for the univariate case (if X has finitely many distinct values).
- Allows more general censoring.
- Each update is a weighted quantile regression problem
  - has to estimate all the quantiles below τ;
  - assumes all the quantile functions are linear in covariates.

Wang and Wang (2009)

- Provides an alternative method to estimate $\tilde{\tau}_i = F(C_i|\mathbf{x}_i)$ by using the **local Kaplan-Meier** estimator (Beran, 1981) of $F(\cdot|\mathbf{x})$.
- No recursive fitting is required.
- Linearity of quantile function is needed only at the quantile level of interest.
- Challenging for high dimensional data.

Peng and Huang (2008)

Idea II: extend the martingale representation of the Nelson-Aalen estimator of the cumulative hazard function to produce an "estimating equation" for conditional quantiles.

- $\Lambda_T(t|\mathbf{x}) = -\log\{1 - F_T(t|\mathbf{x})\}$: cumulative hazard function of T conditional on $\mathbf{x}$;
- $N_i(t) = I(Y_i \leqslant t, \delta_i = 1)$;
- $M_i(t) = N_i(t) - \Lambda_T\{t \wedge Y_i|\mathbf{x}_i)$ is a martingale process so that $E\{M_i(t)|\mathbf{x}_i\} = 0$ for all $t \geqslant 0$.

So

$$E\left[N_i\{\mathbf{x}_i^T\beta_0(\tau)\} - \Lambda_T\{\mathbf{x}_i^T\beta_0(\tau) \wedge Y_i\}|\mathbf{x}_i\right] = 0.$$

Peng and Huang (2008)

Connection between Λ_T and the quantile functions:

$$\begin{aligned}\Lambda_T\{\mathbf{x}_i^T\beta_0(\tau) \wedge Y_i|\mathbf{x}_i\} &= H(\tau) \wedge H\{F_T(Y_i|\mathbf{x}_i)\} \\ &= \int_0^\tau I\{Y_i \geqslant \mathbf{x}_i^T\beta_0(u)\}dH(u),\end{aligned}$$

where $H(u) = -\log(1-u)$ for $0 \leqslant u \leqslant 1$.

The estimating equation becomes

$$n^{-1/2}\sum_{i=1}^n \mathbf{x}_i\left[N_i(\mathbf{x}_i^T\beta) - \int_0^\tau I\{Y_i \geqslant \mathbf{x}_i^T\beta(u)\}dH(u)\right] = 0.$$

Peng and Huang (2008)

Approximating the integral on a grid, $0 = \tau_0 < \tau_1 < \cdots < \tau_J < 1$ yields a simple linear programming formulation to be solved at the gridpoints,

$$\alpha_i(\tau_j) = \sum_{k=0}^{j-1} I\{Y_i \geqslant \mathbf{x}_i^T\hat{\beta}(\tau_k)\}\{H(\tau_{k+1}) - H(\tau_k)\},$$

yielding Peng and Huang's final estimating equation,

$$n^{-1/2}\sum \mathbf{x}_i\left[N_i\{\mathbf{x}_i^T\beta(\tau)\} - \alpha_i(\tau)\right] = 0.$$

Setting $r_i(\mathbf{b}) = Y_i - \mathbf{x}_i^T\mathbf{b}$, this convex function for the Peng and Huang problem takes the form

$$R(\mathbf{b}, \tau_j) = \sum_{i=1}^n r_i(\mathbf{b})\left[\alpha_i(\tau_j) - I\{r_i(\mathbf{b}) < 0\}\delta_i\right] = \min!$$

About Portnoy and Peng-Huang Estimators

- Simulation evidence confirms the asymptotic conclusion that the Portnoy and Peng-Huang estimators are similar (Koenker, 2008; Peng, 2012).
- Both methods require estimation of all the quantiles below τ, and they assume global linearity of quantile functions.

Implementation in R

R syntax:

```
library(quantreg)
#fixed censoring
fit=crq(Curv(y,yc, ctype="left") ~ x, tau = tau,
method = "Pow")
summary(fit)

#Random censoring
fit=crq(Surv(log(time), delta, type="left") ~ x,
method=c("Por", "PengHuang"))
summary(fit)
```


Some One-sample Asymptotics

- Suppose that we have a random sample of pairs, $\{(T_i, C_i) : i = 1, \cdots, n\}$ with $T_i \sim F$, $C_i \sim G$, and T_i and C_i independent.
- Let $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i < C_i)$.
- In this setting the Powell estimator of $\theta = F^{-1}(\tau)$,

$$\hat{\theta}_P = \text{argmin}_\theta \sum_{i=1}^{n} \rho_\tau(Y_i - \min\{\theta, C_i\})$$

is asymptotically normal

$$\sqrt{n}(\hat{\theta}_P - \theta) \rightsquigarrow \mathcal{N}\left(0, \frac{\tau(1-\tau)}{f^2(\theta)\{1 - G(\theta)\}}\right).$$

One-sample Asymptotics

In contrast,

$$\sqrt{n}(\hat{\theta}_{KM} - \theta) \rightsquigarrow \mathcal{N}\left(0, \text{Avar}\left\{\hat{S}(\theta)\right\}/f^2(\theta)\right),$$

$$\text{where Avar}\{\hat{S}(t)\} = S^2(t)\int_0^t \{1 - H(u)\}^{-2} d\tilde{F}(u),$$

$1 - H(u) = \{1 - F(u)\}\{1 - G(u)\}$, $\tilde{F}(u) = \int_0^t \{1 - G(u)\} dF(u)$ and $S(t) = 1 - F(t)$.

The Powell estimator makes use of more sample information. Is it more efficient? The answer is **NO!**

Kaplan Meier v.s. Powell

Proposition

$\text{Avar}(\hat{\theta}_{KM}) \leqslant \text{Avar}(\hat{\theta}_P)$.

Proof.

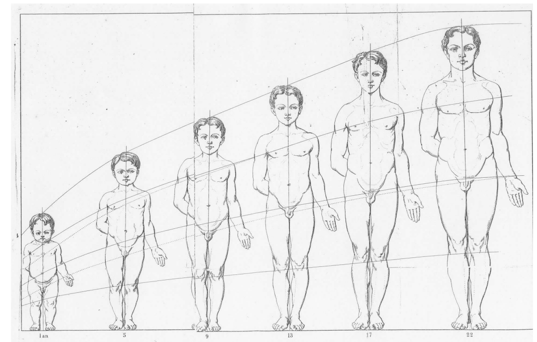
$$\begin{aligned}
f^2(\theta)\text{Avar}(\hat{\theta}_{KM}) &= S(\theta)^2 \int_0^\theta (1 - H(s))^{-2} d\tilde{F}(s) \\
&= S(\theta)^2 \int_0^\theta (1 - G(s))^{-1}(1 - F(s))^{-2} dF(s) \\
&\leqslant \frac{S(\theta)^2}{1 - G(\theta)} \int_0^\theta (1 - F(s))^{-2} dF(s) \\
&= \frac{S(\theta)^2}{1 - G(\theta)} \cdot \frac{1}{1 - F(s)}\Big|_0^\theta \\
&= \frac{S(\theta)^2}{1 - G(\theta)} \cdot \frac{F(\theta)}{1 - F(\theta)} \\
&= \frac{F(\theta)(1 - F(\theta))}{(1 - G(\theta))} \\
&= \frac{\tau(1 - \tau)}{(1 - G(\theta))}.
\end{aligned}$$

□

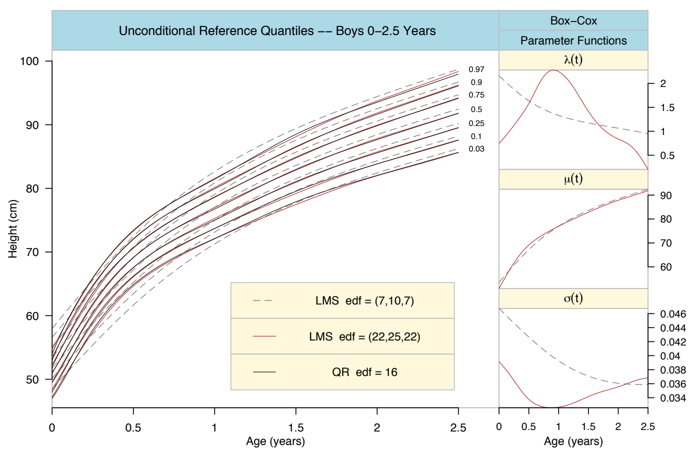
Conclusions

- The Powell estimator, though conceptually attractive, suffers from some serious computational difficulties, imposes strong data requirements, and has a loss in asymptotic efficiency.
- Even for fixed censoring where C_i are observed, it is better to use Portnoy or Peng-Huang estimators than Powell estimator.
- Quantile regression provides a flexible complement to classical survival analysis methods, and is now well equipped to handle censoring.

Lecture 5. Nonparametric Quantile Regression



In the Beginning, . . . were the Quantiles



Wei et al. (2006)

Three Approaches to Nonparametric QR

- Locally Polynomial (Kernel) Methods: `lprq`

$$\hat{\alpha}(\tau, x) = \text{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha_0 - \alpha_1(x_i - x) - \ldots - \frac{1}{p!}\alpha_p(x_i - x)^p)$$
$$\hat{g}(\tau, x) = \hat{\alpha}_0(\tau, x)$$

- Series Methods `rq( y ~ bs(x,knots = k) + z)`

$$\hat{\alpha}(\tau) = \text{argmin}_\alpha \sum_{i=1}^{n} \rho_\tau(y_i - \sum_j \varphi_j(x_i)\alpha_j)$$
$$\hat{g}(\tau, x) = \sum_{j=1}^{p} \varphi_j(x)\hat{\alpha}_j$$

- Penalty Methods `rqss`

$$\hat{g}(\tau, x) = \text{argmin}_g \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda P(g)$$

Total Variation Regularization I

There are many possible penalties, ways to measure the roughness of fitted function, but total variation of the first derivative of g is particularly attractive:

$$P(g) = V(g') = \int |g''(x)|dx.$$

As $\lambda \to \infty$ we constrain g to be closer to linear in x. Solutions of

$$min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda V(g')$$

are continuous and piecewise linear.

Fish in a Bottle

Objective: to study metabolic activity of various fish species in an effort to better understand the nature of the feeding cycle. Metabolic rates based on oxygen consumption as measured by sensors mounted on the tubes.

Three primary aspects are of interest:

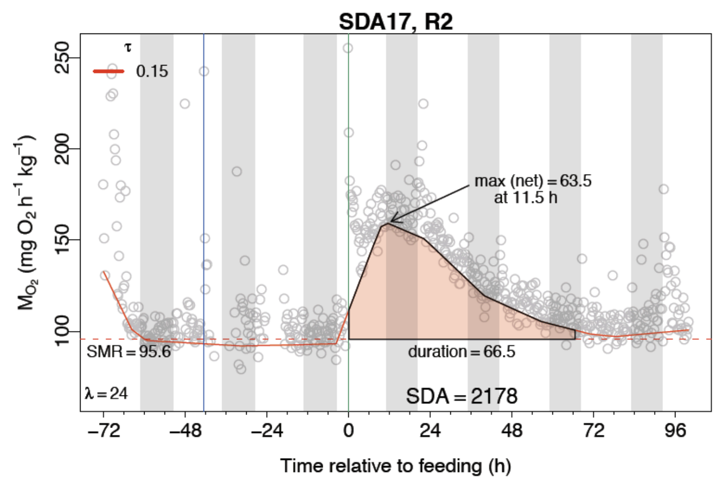
1. Basal (minimal) Metabolic Rate,
2. Duration and Shape of the Feeding Cycle, and
3. Diurnal Cycle.

Some Experimental Details

Experimental data of Denis Chabot, Institut Maurice-Lamontagne, Quebec, Canada and his colleagues.

1. Basal (minimal) metabolic rate M_{O_2} (aka Standard Metabolic Rate SMR) is measured in $mg\ O_2\ h^{-1}\ kg^{-1}$ for fish "at rest" after several days without feeding,
2. Fish are then fed and oxygen consumption monitored until M_{O_2} returns to its prior SMR level for several hours.
3. Elevation of M_{O_2} after feeding (aka Specific Dynamic Action SDA) ideally measures the energy required for digestion,
4. Procedure is repeated for several cycles, so each estimation of the cycle is based on a few hundred observations.

Juvenile Codfish



Tuning Parameter Selection

There are two tuning parameters:

1. $\tau = 0.15$ the (low) quantile chosen to represent the SMR,
2. λ controls the smoothness of the SDA cycle.

One way to interpret the parameter λ is to note that it controls the number of effective parameters of the fitted model (Meyer and Woodroofe, 2000):

$$p(\lambda) = \text{div } \hat{g}_{\lambda,\tau}(y_1, ..., y_n) = \sum_{i=1}^{n} \partial \hat{y}_i / \partial y_i$$

This is equivalent to the number of interpolated observations, the number of zero residuals. Selection of λ can be made by minimizing, e.g. Schwarz Criterion:

$$\text{SIC}(\lambda) = n \log(n^{-1} \sum \rho_\tau(y_i - \hat{g}_{\lambda,\tau}(x_i))) + \frac{1}{2} p(\lambda) \log n.$$

Total Variation Regularization II

For bivariate functions we consider the analogous problem:

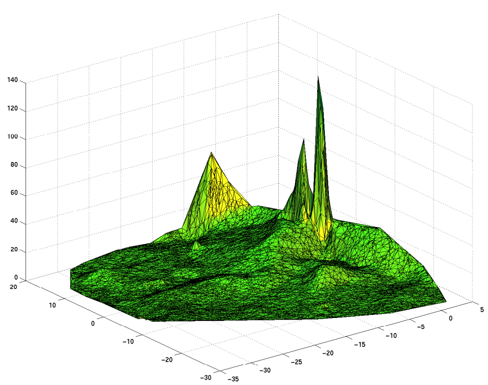
$$min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_{1i}, x_{2i})) + \lambda V(\nabla g),$$

where the total variation variation penalty is now:

$$V(\nabla g) = \int \|\nabla^2 g(x)\| dx.$$

Solutions are again continuous, but now they are piecewise linear on a triangulation of the observed x observations. Again, as $\lambda \to \infty$ solutions are forced toward linearity.

Chicago Land Values via TV Regularization



Chicago Land Values: Based on 1194 vacant land sales and 7505 "virtual" sales introduced to increase the flexibility of the triangulation.

Additive Models: Putting the pieces together

We can combine such models:

$$min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - \sum_j g_j(x_{ij})) + \sum_j \lambda_j V(\nabla g_j)$$

- Components g_j can be univariate, or bivariate.
- Additivity is intended to muffle the curse of dimensionality.
- Linear terms are easily allowed, or enforced.
- And shape restrictions like monotonicity and convexity/concavity as well as boundry conditions on g_j's can also be imposed.

Implementation in the R quantreg Package

- Problems are typically large, very sparse linear programs.
- Optimization via interior point methods are quite efficient,
- Provided sparsity of the linear algebra is exploited, quite large problems can be estimated.
- The nonparametric qss components can be either univariate, or bivariate
- Each qss component has its own λ specified
- Linear covariate terms enter formula in the usual way
- The qss components can be shape constrained.

```
fit <- rqss(y ~ qss(x1,3) + qss(x2,8) + x3, tau =
                     .6)
```


Pointwise Confidence Bands

It is obviously crucial to have reliable confidence bands for nonparametric components. Following Wahba (1983) and Nychka(1983), conditioning on the λ selection, we can construct bands from the covariance matrix of the full model:

$$V = \tau(1-\tau)(\tilde{X}^\top \Psi \tilde{X})^{-1}(\tilde{X}^\top \tilde{X})^{-1}(\tilde{X}^\top \Psi \tilde{X})^{-1}$$

with

$$\tilde{X} = \begin{bmatrix} X & G_1 & \cdots & G_J \\ \lambda_0 H_K & 0 & \cdots & 0 \\ 0 & \lambda_1 P_1 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_j P_J \end{bmatrix} \quad \text{and} \quad \Psi = \text{diag}(\phi(\hat{u}_i/h_n)/h_n)$$

Pointwise bands can be constructed by extracting diagonal blocks of V.

Uniform Confidence Bands

Uniform bands are also important, but more challenging. We would like:

$$B_n(x) = (\hat{g}_n(x) - c_\alpha \hat{\sigma}_n(x), \hat{g}_n(x) + c_\alpha \hat{\sigma}_n(x))$$

such that the true curve, g_0, is covered with specified probability $1-\alpha$ over a given domain $\mathcal{X}$:

$$\mathcal{P}\{g_0(x) \in B_n(x) \mid x \in \mathcal{X}\} \geqslant 1-\alpha.$$

We can follow the “Hotelling tube” approach based on Hotelling(1939) and Weyl (1939) as developed by Naiman (1986), Johansen and Johnstone (1990), Sun and Loader (1994) and others.

Uniform Confidence Bands

Hotelling’s original formulation for parametric nonlinear regression has been extended to non-parametric regression. For series estimators

$$\hat{g}_n(x) = \sum_{j=1}^{p} \varphi_j(x)\hat{\theta}_j$$

with pointwise standard error $\sigma(x) = \sqrt{\varphi(x)^\top V^{-1} \varphi(x)}$ we would like to invert test statistics of the form:

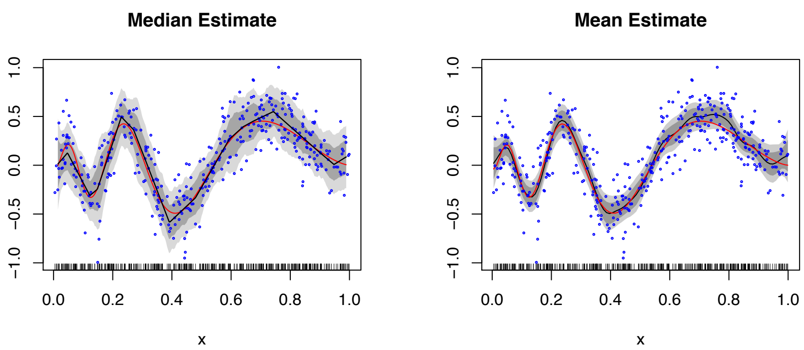
$$T_n = \sup_{x \in \mathcal{X}} \frac{\hat{g}_n(x) - g_0(x)}{\sigma(x)}.$$

This requires solving for the critical value, c_α in

$$\mathcal{P}(T_n > c) \leqslant \frac{\kappa}{2\pi}(1 + c^2/\nu)^{-\nu/2} + \mathcal{P}(t_\nu > c) = \alpha$$

where κ is the length of a “tube” determined by the basis expansion, t_ν is a Student random variable with degrees of freedom $\nu = n - p$.

Confidence Bands in Simulations



$$Y_i = \sqrt{x_i(1-x_i)}\,\sin\left(\frac{2\pi(1+2^{-7/5})}{x_i+2^{-7/5}}\right) + U_i, \quad i = 1, \cdots, 400, \quad U_i \sim \mathcal{N}(0, 0.04)$$

,

Simulation Performance

	Accuracy			Pointwise		Uniform	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.063	0.046	12.936	0.960	0.999	0.323	0.920
gam	0.045	0.035	20.461	0.956	0.998	0.205	0.898
t_3							
rqss	0.071	0.052	11.379	0.955	0.998	0.274	0.929
gam	0.071	0.054	17.118	0.948	0.994	0.159	0.795
t_1							
rqss	0.099	0.070	9.004	0.930	0.996	0.161	0.867
gam	35.551	2.035	8.391	0.920	0.926	0.203	0.546
χ^2_3							
rqss	0.110	0.083	8.898	0.950	0.997	0.270	0.883
gam	0.096	0.074	14.760	0.947	0.987	0.218	0.683

Performance of Penalized Estimators and Their Confidence Bands: IID Error Model

Simulation Performance

	Accuracy			Pointwise		Uniform	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.081	0.063	10.685	0.951	0.998	0.265	0.936
gam	0.064	0.050	17.905	0.957	0.999	0.234	0.940
t_3							
rqss	0.091	0.070	9.612	0.952	0.998	0.241	0.938
gam	0.103	0.078	14.656	0.949	0.992	0.232	0.804
t_1							
rqss	0.122	0.091	7.896	0.938	0.997	0.222	0.893
gam	78.693	4.459	7.801	0.927	0.958	0.251	0.695
χ^2_3							
rqss	0.145	0.114	7.593	0.947	0.998	0.307	0.921
gam	0.138	0.108	12.401	0.941	0.973	0.221	0.626

Performance of Penalized Estimators and Their Confidence Bands: Linear Scale Model

Example 3: Childhood Malnutrition in India

A larger scale problem illustrating the use of these methods is a model of risk factors for childhood malnutrition considered by Fenske et al. (2011).

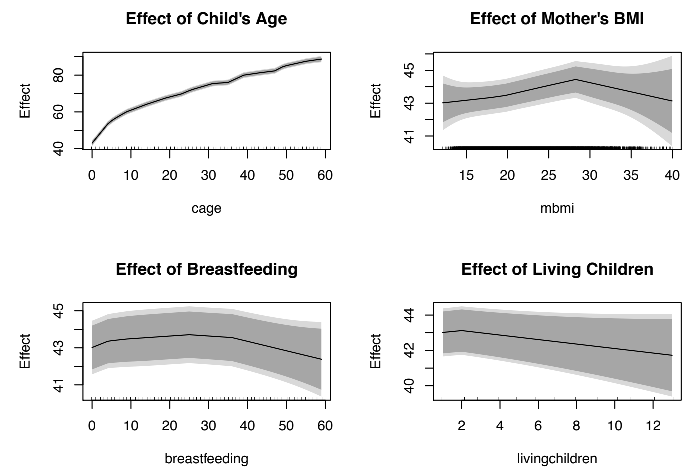
- They motivate the use of models for low conditional quantiles of height as a way to explore influences on malnutrition,
- They employ boosting as a model selection device,
- Their model includes six univariate nonparametric components and 15 other linear covariates.
- There are 37,623 observations on the height of children from India.

Example 3: R Formulation

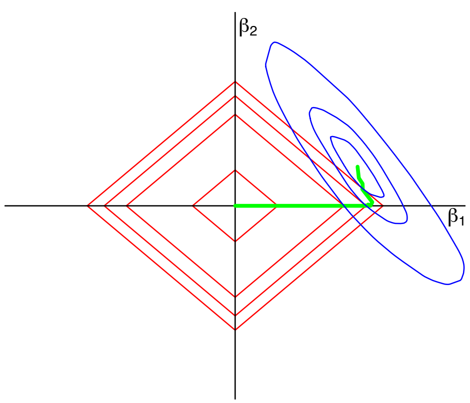
```
 fit <- rqss(cheight ~ qss(cage, lambda = lam[1]) +
qss(bfed, lambda = lam[2]) + qss(mage, lambda = lam[3]) +
qss(mbmi, lambda = lam[4]) + qss(sibs, lambda = lam[5]) +
qss(medu, lambda = lam[6]) + qss(fedu, lambda = lam[7]) +
csex + ctwin + cbirthorder + munemployed + mreligion +
mresidence + deadchildren + wealth + electricity +
radio + television + frig + bicycle + motorcycle + car +
tau = 0.10, method = "lasso", lambda = lambda, data = india)
```

- The seven coordinates of `lam` control the smoothness of the nonparametric components,
- `lambda` controls the degree of shrinkage in the linear (lasso) coefficients.
- The estimated model has roughly 40,000 observations, including the penalty contribution, and has 2201 parameters.
- Fitting the model for a single choice of λ's takes approximately 5 seconds.

Example 3: Selected Smooth Components



Example 3: Lasso Shrinkage of Linear Components



Lasso λ Selection – Another Approach

Lasso shrinkage is a special form of the TV penalty:

$$R_\tau(b) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b)$$

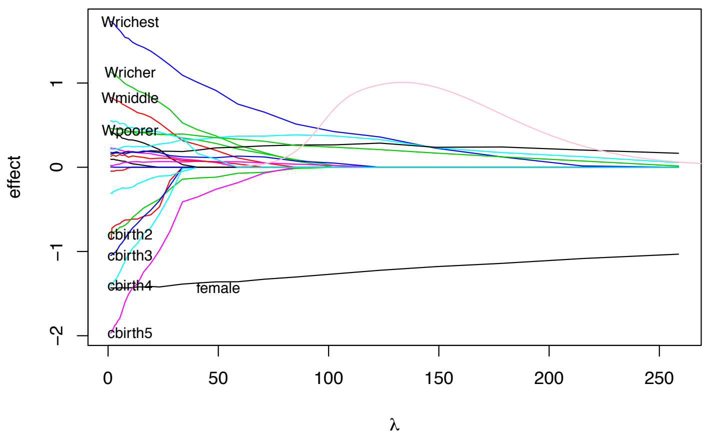
$$\begin{aligned} \hat{\beta}_{\tau,\lambda} &= \text{argmin}\{R_\tau(b) + \lambda \|b\|_1\} \\ &\in \{b : \ 0 \in \partial R_\tau(b) + \lambda \partial \|b\|_1\}. \end{aligned}$$

At the true parameter, $\beta_0(\tau)$, we have the pivotal statistic,

$$\begin{aligned} \partial R_\tau(\beta_0(\tau)) &= \sum (\tau - I(F_{y_i}(y_i) \leqslant \tau)) x_i \\ &\sim \sum (\tau - I(U_i \leqslant \tau)) x_i \end{aligned}$$

Proposal: (Belloni and Chernozhukov, 2009) Choose λ as the $1 - \alpha$ quantile of the simulated distribution of $\| \sum (\tau - I(U_i \leqslant \tau)) x_i \|_\infty$ with iid $U_i \sim U[0, 1]$.

Example 3: Lasso Shrinkage of Linear Components



Conclusions

- Nonparametric specifications of $Q(\tau|x)$ improve flexibility.
- Additive models keep effective dimension in check.
- Total variation roughness penalties are natural.
- Schwarz model selection criteria are useful for λ selection.
- Hotelling tubes are useful for uniform confidence bands.
- Lasso Shrinkage is useful for parametric components.

Lecture 6. Bayesian Quantile Regression

- Introduction
- Bayesian QR Based on Asymmetric Laplace Likelihood
- Bayesian Empirical Likelihood
- Nonparametric/Semiparametric Likelihood
  - Mixtures with Dirichlet Process Priors
  - Semiparametric Bayesian for Simultaneous Linear Quantile Regression
  - Approximate Likelihood

Introduction

- Linear quantile regression model:

$$Q_{y_i}(\tau|\mathbf{x}_i) = \mathbf{x}_i^T\beta(\tau), i = 1, \ldots, n.$$

- Frequentist estimator of $\beta(\tau)$:

$$\hat{\beta}(\tau) = \text{argmin}_\beta \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\beta), \quad \rho_\tau(u) = u\{\tau - I(u < 0)\}.$$

- Advantages of Bayesian QR:
  - point estimates and confidence intervals can be calculated simultaneously from the posterior sequences;
  - the use of MCMC can help avoid the suffering from the computational curse of dimensionality, and the difficulty in the optimization of a (highly) nonconvex objective function such as for Powell's estimator for fixed censored data.

Introduction

- Bayesian QR is challenging as quantile regression typically does not assume any parametric likelihood.
- A working likelihood is needed for Bayesian QR:
  - parametric working likelihood, e.g. asymmetric Laplacian (Yu and Moyeed, 2001; Geraci and Bottai, 2007);
  - nonparametric/semiparametric working likelihood, e.g. Gelfand and Kottas (2002); Kottas and Krnjajić (2009); Reich et al. (2010); Dunson and Taylor (2005); Reich et al. (2011);
  - empirical likelihood (Lancaster and Jun, 2010; Otsu, 2008; Yang and He, 2012).

Asymmetric Laplacian (AL) Likelihood

A random variable Z is said to follow an asymmetric Laplace distribution $AL(\mu, \sigma, \tau)$ if its density is given by

$$f(z) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{z-\mu}{\sigma}\right)\right\}.$$

Therefore, the **MLE of** μ (assume τ is known):

$$\text{argmin}_\mu \sum_{i=1}^n \rho_\tau\left(\frac{z_i-\mu}{\sigma}\right) = \text{argmin}_\mu \sum_{i=1}^n \rho_\tau(z_i-\mu)$$

is just the **sample quantile** of $(z_1, \cdots, z_n)$.

Asymmetric Laplacian (AL) Likelihood

Yu and Moyeed (2001) developed a Bayesian quantile regression method assuming **AL likelihood** for $\mathbf{y} = (y_1, \cdots, y_n)$:

$$L(\mathbf{y}|\boldsymbol{\beta}) = \{\tau(1-\tau)\}^n \exp\left\{-\sum_{i=1}^n \rho_\tau\{y_i - \mathbf{x}_i^T\boldsymbol{\beta})\right\},$$

that is, assuming $Y|\mathbf{x}_i \sim AL(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \tau)$.

Posterior distribution of $\boldsymbol{\beta} = \boldsymbol{\beta}(\tau)$:

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}),$$

where $\pi(\boldsymbol{\beta})$ is the prior distribution of $\boldsymbol{\beta}$.

Properties of the AL-based Bayesian QR:

- If the flat prior $\pi(\boldsymbol{\beta}) \propto 1$ is used, then
  - the posterior distribution of $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta}|\mathbf{y})$ is proper;
  - the posterior mode is the frequentist estimator $\hat{\boldsymbol{\beta}}(\tau)$.
- However, when the AL likelihood is misspecified,
  - posterior chain from the Bayesian AL quantile regression **does not lead to valid posterior inference**;
  - correction to the covariance matrix of the posterior chain is possible to enable an asymptotically valid posterior inference (Chernozhukov and Hong, 2003; Yang et al., 2014).

Other Bayesian AL QR Developments

Geraci and Bottai (2007), Geraci and Bottai (2013)

- Quantile regression with a random intercept effect:

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, b_i) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i.$$

- Assume $(Y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\eta}, b_i) \sim AL(\mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i, \sigma, \tau)$ and $b_i \sim N(0, \varphi^2)$, where $\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma, \varphi)$.
- Estimate $\boldsymbol{\eta}$ with EM by integrating out b_i from $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\eta}) = f(y|\boldsymbol{\eta}, \mathbf{b})f(\mathbf{b}|\boldsymbol{\eta})$.

Other Bayesian AL QR Developments

Tsionas (2003), Kozumi and Kobayashi (2011): developed Gibbs sampling procedures using the conditionally Gaussian representation for AL:

if $\epsilon \sim AL(0, \sigma, \tau)$, then we can represent ϵ as:

$$\epsilon = \sqrt{\frac{2\xi\sigma}{\tau(1-\tau)}}Z + \frac{(1-2\tau)}{\tau(1-\tau)}\xi,$$

where $Z \sim N(0,1)$, and $\xi \sim Gamma(1, 1/\sigma) = Exp(1/\sigma)$.

Other Bayesian AL QR Developments

MCMC algorithm for Bayesian linear regression with Laplace errors (Choi and Hobert, 2013).

- Model: $Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma\epsilon_i$, $\epsilon_i \sim Laplace(0,1) = AL(0,1,1/2)$.
- Data augmentation algorithm: based on the representation of the Laplace density as a scale mixture of normals with respect to the inverse gamma distribution, i.e.

$$Z \sim IG(1, 1/8), Y|Z = z \sim N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2/z)$$

$$\Rightarrow Y|\mathbf{x} \sim Laplace(\mathbf{x}^T\boldsymbol{\beta}, \sigma).$$

- Choi and Hobert (2013) showed that the MCMC underlying the DA algorithm is geometrically ergodic, which guarantees the existence of the central limit theorems that form the basis of all the standard methods of calculating valid asymptotic standard errors for MCMC-based estimators.

Other Bayesian AL QR Developments

- Li et al. (2010): Bayesian regularized quantile regression.
- Lum and Gelfand (2012): Bayesian spatial quantile regression assuming asymmetric Laplace process:

$$\begin{aligned} Y(s) &= \mu_\tau(s) + \epsilon_\tau(s), \quad \mu_\tau(s) = \mathbf{x}^T(s)\boldsymbol{\beta}_\tau, \\ \epsilon_\tau(s) &= \sqrt{\frac{2\xi(s)\sigma}{\tau(1-\tau)}}Z(s) + \frac{1-2\tau}{\tau(1-\tau)}\xi(s), \end{aligned}$$

where $Z(s)$ is a Gaussian process, and $\xi(s)$ is a process with marginal to be exponential with rate $1/\sigma$.

Bayesian Empirical Likelihood 1

Suppose we observe a random sample $(y_i, \mathbf{x}_i)$ from the quantile regression model

$$Q_\tau(Y|\mathbf{X}) = \mathbf{X}^T\beta(\tau), \qquad (6)$$

Conventional estimator of β:

$$\widehat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T\beta),$$

which is also a solution to the estimating equation

$$n^{-1/2}\sum_{i=1}^n \mathbf{x}_i\psi_\tau(y_i - \mathbf{x}_i^T\beta) \approx 0,$$

where $\psi_\tau(u) = \tau - I(u < 0)$.

Bayesian Empirical Likelihood 2

For a proposed β, profile empirical likelihood (EL):

$$\begin{aligned} R(\beta) &= \max\Big\{\prod_{i=1}^n p_i \Big| \sum_{i=1}^n p_i \boldsymbol{m}_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta}) = 0, \\ &\qquad \sum_{i=1}^n p_i = 1,\ 0 \leqslant p_i \leqslant 1\Big\}, \end{aligned}$$

where p_i is the weight for the ith observation, and $\boldsymbol{m}_\tau(\mathbf{x}_i, y_i, \beta) = \mathbf{x}_i\psi_\tau(y_i - \mathbf{x}_i^T\beta)$ is the score function.

Bayesian Empirical Likelihood 3

For a given β, the EL $L_n(\beta)$ can be calculated by using the standard Lagrange multiplier method (Owen, 2001):

$$\begin{aligned} R(\boldsymbol{\beta}) &= \prod_i p_i(\boldsymbol{\beta}), \\ p_i(\boldsymbol{\beta}) &= [n\{1 + \lambda(\boldsymbol{\beta})\boldsymbol{m}_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta})\}]^{-1}, \end{aligned}$$

where $\lambda(\boldsymbol{\beta})$ satisfies

$$\sum_i \frac{\boldsymbol{m}_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta})}{1 + \lambda(\boldsymbol{\beta})\boldsymbol{m}_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta})} = 0.$$

BEL posterior density:

$$f(\boldsymbol{\beta}|D) \propto R(\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta}).$$

Bayesian Empirical Likelihood 4

Multiple quantiles

- Consider the quantile regression model at $\tau_1 < \cdots < \tau_k$, k quantile levels.
- Stacked estimating function

  $$\boldsymbol{m}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = (\boldsymbol{m}_{\tau_1}^T(\mathbf{x}_i, y_i, \boldsymbol{\theta}), \cdots, \boldsymbol{m}_{\tau_k}^T(\mathbf{x}_i, y_i, \boldsymbol{\theta}))^T,$$

  where $\boldsymbol{\theta} = (\boldsymbol{\beta}(\tau_1)^T, \cdots, \boldsymbol{\beta}(\tau_k)^T)^T$.
- The similar procedure follows by replacing $\boldsymbol{m}_\tau$ by $\boldsymbol{m}$, $\boldsymbol{\beta}$ by $\boldsymbol{\theta}$, and $\lambda(\boldsymbol{\beta})m_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta})$ by $\boldsymbol{\lambda}(\boldsymbol{\theta})^T\boldsymbol{m}(\mathbf{x}_i, y_i, \boldsymbol{\theta})$.

Bayesian Empirical Likelihood 5

Properties

- Asymptotically the posterior variance is the same as that of the sampling variance of quantile regression estimator.
- The method allows joint modeling of multiple quantiles to borrow information across quantiles to improve efficiency.
- The empirical likelihood can be generalized to a class of nonparametric likelihoods, such as exponential tilting.

Bayesian Empirical Likelihood 6

Some remarks

- The BEL posterior is not really a posterior.
- No need to solve for the maximum EL estimator, and the calculation of $R(\boldsymbol{\theta})$ is easy.
- Metropolis-Hastings algorithm is used for sampling from the posterior.
- An improper prior cannot guarantee a proper posterior.
- The posterior will be improper for flat priors on $\boldsymbol{\theta}$.
- Shrinking priors on $\boldsymbol{\theta}$ can be used to shrink quantile slopes towards common values or a pre-specified parametric form.

Reference: Yang and He (2012).

Nonparametric/Semiparametric Likelihood

Mixtures with Dirichlet Process Priors

- Suppose $\epsilon \sim AL(0, \tau, \sigma)$ with density

$$k_\tau^{AL}(\epsilon;\sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{\epsilon}{\sigma}\right)\right\},$$

  where $\sigma > 0$ is a scale parameter and $\tau \in (0,1)$. Then ϵ has the τth quantile zero, i.e. $\int_{-\infty}^{0} k_p^{AL}(\epsilon;\sigma) d\epsilon = \tau$.
- Thus mixture of $k_\tau^{AL}(\epsilon;\sigma_i), i = 1, \cdots, n$ also has the τth quantile zero.

(A). Kottas and Krnjajić (2009)

Model 1: scale mixture of AL densities

$$\begin{aligned} &Y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma_i \overset{ind.}{\sim} k_\tau^{AL}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}; \sigma_i), i = 1, \cdots, n, \\ &\sigma_i|G \overset{i.i.d.}{\sim} G, i = 1, \cdots, n, \\ &G|\alpha, d \sim \text{Drichilet process } DP(\alpha, G_0), \end{aligned}$$

This mixture model

- preserves the zero τth quantile for the residual;
- is more flexible than the AL distribution;
- it has the same issue as the AL likelihood approach, both have lack of coherence for multiple quantiles.

Kottas and Krnjajić (2009)

Model 2: nonparametric scale mixture of uniform densities

$$k_\tau(\epsilon;\sigma_1,\sigma_2) = \frac{\tau}{\sigma_1} I(-\sigma_1 < \epsilon < 0) + \frac{1-\tau}{\sigma_2} I(0 \leqslant \epsilon < \sigma_2), \sigma_r > 0, r = 1, 2.$$

has the τth quantile zero. Hierarchical model:

$$\begin{aligned} &Y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma_{1i}, \sigma_{2i} \overset{ind.}{\sim} k_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}; \sigma_{1i}, \sigma_{2i}), i = 1, \cdots, n, \\ &\sigma_{ri}|G_r \overset{i.i.d.}{\sim} G_r, r = 1, 2, i = 1, \cdots, n, \\ &G_r|\alpha_r, d_r \sim DP(\alpha_r, G_{r0}), r = 1, 2. \end{aligned}$$

- Model 2 poses separate priors for the negative and positive residuals to ensure zero τth quantiles.
- The resulting conditional distribution of Y is discontinuous at the mode.
- The model forces the mode to be at the quantile of interest.

(B). Reich et al. (2010)

Assume the location-scale shift model

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{x}_i^T\gamma\epsilon_i,$$

where $\mathbf{x}_i^T\gamma > 0$ for all $\mathbf{x}_i$, and ϵ_i are $i.i.d.$ with τth quantile zero. Assume ϵ_i follows the infinite mixture distribution

$$h(\epsilon|\mu,\sigma^2) = \sum_{k=1}^{\infty} p_k f(\epsilon|\mu_k, \sigma_k^2, q_k), \ \sum_{k=1}^{\infty} p_k = 1,$$

- $f(\epsilon|\mu_k, \sigma_k^2, q_k)$ is the quantile-restricted 2-component mixture
$$f(\epsilon|\mu_k,\sigma_k^2,q_k) = q_k\phi(\mu_{1k},\sigma_{1k}^2) + (1-q_k)\phi(\mu_{2k},\sigma_{2k}^2);$$
- to ensure that $\int_{-\infty}^{0} f(\epsilon|\mu_k,\sigma_k^2,q_k)d\epsilon = \tau$, let
$$q_k = \frac{\tau - \Phi(-\mu_{2k})/\sigma_{2k}}{\Phi(-\mu_{1k}/\sigma_{1k}) - \Phi(-\mu_{2k}/\sigma_{2k})}.$$

Tokdar and Kadane (2011)

Tokdar and Kadane (2011): **Semiparametric Bayesian for Simultaneous Linear Quantile Regression**

For an univariate x:

- $Q_\tau(Y|x) = \beta_0(\tau) + \beta_1(\tau)x, \tau \in [0,1]$ is monotonically increasing in τ for every $x \in [-1,1]$ if and only if
$$Q_\tau(Y|x) = \mu + \gamma x + \frac{1-x}{2}\eta_1(\tau) + \frac{1+x}{2}\eta_2(\tau),$$
where $\eta_1(\tau)$ and $\eta_2(\tau)$ are monotonically increasing in $\tau \in [0,1]$.

Tokdar and Kadane (2011)

- Pose a prior for $\eta_1(\tau)$ and $\eta_2(\tau)$ by a logistic transformation of a smooth Gaussian process:
  - $\eta_1(\tau) = \sigma_1\tilde{Q}\{\xi_1(\tau)\}$;
  - $\eta_2(\tau) = \sigma_2\tilde{Q}\{\xi_2(\tau)\}$;
  - $\tilde{Q}$ is the quantile function of a target parametric distribution whose support is (y_{min}, y_{max});
  - $\xi_j(\tau), j = 1, 2$ are monotonically increasing random functions mapping $[0,1]$ to $[0,1]$, defined through a logistic transformation of a smooth Gaussian process.
- Log-likelihood function
$$\sum_i \log f_Y(y_i|x_i) = -\sum_i \log\left[\frac{\partial}{\partial\tau}Q_\tau(Y|x_i)\Big|_{\tau_i}\right],$$
where $\tau_i = \{\tau : y_i = Q_\tau(Y|x_i)\}$.

Approximate Likelihood (Reich et al., 2011)

Reich et al. (2011): used an approximate likelihood from the multivariate normal distribution of the joint quantile regression estimates.

Main ideas:

- Assume a global linear quantile regression model:

$$Q_\tau(Y|\mathbf{x}) = \mathbf{x}^T\beta(\tau), \tau \in (0,1).$$

- Using the observed data $\{(y_i, \mathbf{x}_i), i = 1, \cdots, n\}$, fit quantile regression at each quantile level from $0 < \tau_1 < \cdots < \tau_K < 1$ separately to obtain $\hat{\beta}(\tau_k), k = 1, \cdots, K$.
- Fit the approximate model

$$(\hat{\beta}(\tau_k), k = 1, \cdots, K) \sim N\big(\beta(\boldsymbol{\tau}), \Sigma\big),$$

where $\beta(\boldsymbol{\tau}) = (\beta(\tau_k), k = 1, \cdots, K)$, Σ is the covariance matrix of $(\hat{\beta}(\tau_k), k = 1, \cdots, K)$ and can be estimated using existing methods in R package *quantreg*.

Approximate Likelihood (Reich et al., 2011)

- Model the quantile coefficient process $\beta_j(\tau)$, the jth element of $\beta(\tau)$, by using Bernstein basis polynomials

$$\beta_j(\tau) = \sum_{m=1}^{M} B_m(\tau)\alpha_{jm},$$

where

- $B_m(\tau) = \binom{M}{m}\tau^m(1-\tau)^{M-m}$ are the Bernstein basis polynomials.
- α_{jm} are unknown coefficients with specific priors to ensure monotonicity of the conditional quantile functions.

Lecture 7. Quantile Regression for Longitudinal Data

- Marginal Quantile Regression Model
  - Estimator Based on Working Independence
  - Jung's Estimator
- Conditional Quantile Regression Model
  - Penalized Method–Koenker (2004)
  - Correlated Random Effects
  - Empirical Likelihood

Longitudinal/Clustered Data

- Observed data: $\{(y_{ij}, \mathbf{x}_{ij}), j = 1, \cdots, m_i\}$ are associated with subject i, $i = 1, \cdots, n$. Observations from the same subject tend to be correlated.
- Main challenges for quantile regression
  - quantiles are not linear operators, that is, $Q_{X+Y}(\tau) \neq Q_X(\tau) + Q_Y(\tau)$ in general;
  - usually no parametric likelihood is assumed.
- Two types of models
  - Marginal model
  - Conditional model

Marginal Quantile Regression Model

- Captures the average trend among all subjects.
- Association among repeated measures within subjects: accounted for by modeling the intra-subject correlation among residuals, after removing the average trend.
- A general marginal (linear) quantile regression model

$$Y_{ij} = \mathbf{x}_{ij}^T \beta(\tau) + u_{ij}(\tau), i = 1, \cdots, n, j = 1, \cdots, m_i,$$

  where the τth conditional quantile of $u_{ij}(\tau)$ given $\mathbf{x}_{ij}$ is zero, $u_{ij}(\tau)$ are correlated within the same subject, and independent between subjects.
- Under this model, the τth conditional quantile of Y_{ij} given $\mathbf{x}_{ij}$ is

$$Q_{Y_{ij}}(\tau|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \beta(\tau).$$

Estimation under Working Independence

- Estimate $\beta(\tau)$ pretending that data are independent,

$$\hat{\beta}(\tau) = \text{argmin}_{\mathbf{b} \in R^p} \sum_{ij} \rho_\tau(y_{ij} - \mathbf{x}_{ij}^T \mathbf{b}).$$

- The working independence estimator is still consistent and asymptotically normally distributed

$$n^{1/2}(\hat{\beta}(\tau) - \beta(\tau)) = AN(0, H_n^{-1} J_n H_n^{-1}),$$

  - $H_n = n^{-1} \sum_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T f_{ij}(0)$, f is the density of $u_{ij}(\tau)$,
  - $J_n = n^{-1} \sum_i X_i^T Cov\{\psi_\tau(u_{i1}(\tau)), \cdots, \psi_\tau(u_{im_i}(\tau)\} \mathbf{X}_i$,
  - $\psi_\tau(u) = \tau - I(u < 0)$,
  - $\mathbf{X}_i = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{im_i}^T)^T$.

Inference

To carry out inference for estimators obtained under working independence, correlation needs to be accounted for in the statistical inference.

- Wald test: estimate H and J directly.
- Rank score test: Wang and He (2007).
- Block bootstrap: resample subjects.

More Efficient Estimation

- The estimator obtained under working independence is not most efficient.
- Efficiency can be improved by **incorporating the intra-subject correlation in the estimation** process. But empirical studies show that the improvement is limited except when the intra-subject correlation is extremely high.

Jung's Estimator (Jung, 1996)

Extension of Jung's estimator to quantile regression with clustered data: estimate $\beta(\tau)$ by solving the following estimating equation

$$n^{-1/2}\sum_{i=1}^{n} \mathbf{X}_i^T W_i V_i^{-1}\psi_\tau\{\boldsymbol{Y}_i - \mathbf{X}_i^T\beta(\tau)\} \approx 0, \tag{7}$$

where

- $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{im_i})^T$,
- $W_i = diag\{f_{i1}(0), \cdots, f_{im_i}(0)\}$,
- $f_{ij}(\cdot)$ is the density of $u_{ij}(\tau)$,
- $V_i = Cov\Big(\psi_\tau(y_{i1} - x_{i1}^T\beta(\tau)), \cdots, \psi_\tau(y_{im} - x_{im_i}^T\beta(\tau))\Big)$.

Jung's Estimator (Jung, 1996)

Following the reweighted least squares algorithm, we can compute $\hat\beta$ by iterating

$$\hat\beta \leftarrow \left[\sum_{i=1}^{n} \mathbf{x}_i^T W_i V_i^{-1} A_i \mathbf{x}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^T W_i V_i^{-1} A_i \boldsymbol{Y}_i\right],$$

where $A_i = diag\left\{\psi_\tau(y_{ij} - \mathbf{x}_{ij}^T\hat\beta)/(y_{ij} - \mathbf{x}_{ij}^T\hat\beta)\right\}$ with the convention that $\psi_\tau(u)/u = 0$ when $u = 0$.

Alternatively, we can adopt the induced smoothing method in Brown and Wang (2005) to estimate a smoothed estimator, which has the same asymptotic distribution as $\hat\beta$; see e.g. Leng and Zhang (2014).

Conditional Quantile Regression Model

Conditional model, or subject-specific model

- The model captures an **individual trend** for each subject/cluster.
- Variability between subjects is due to variation in individual trends.
- Association among repeated measures within subjects: arises because all observations on the same subject have the same underlying trend.

Random Intercept Model

- Model:
  $$Y_{ij} = \mathbf{x}_{ij}^T\beta + \mathbf{b_i} + e_{ij},$$
  where b_i is the random subject effect, e_{ij} is the independent random measurement error, and b_i and e_{ij} are independent.
- Under this model: $E(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\beta + E(b_i|\mathbf{x}_{ij}) + E(e_{ij}|\mathbf{x}_{ij})$.
- However, $Q_{b_i+e_{ij}}(\tau|\mathbf{x}_{ij}) \neq Q_{b_i}(\tau|\mathbf{x}_{ij}) + Q_{e_{ij}}(\tau|\mathbf{x}_{ij})$.
- Conditional QR model with a random intercept effect:
  $$Q_{Y_{ij}}(\tau|\mathbf{x}_{ij}, b_i) = \mathbf{x}_{ij}^T\beta(\tau) + b_i, \tag{8}$$
  so each subject has a subject-specific location parameter.

Penalized Method–Koenker (2004)

- Assume the QR model with random intercept effects:

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, b_i) = b_i + \mathbf{x}_{ij}^T\beta(\tau), i = 1, \ldots, n, j = 1, \ldots, m_i.$$

- b_i has purely a location shift effect so it is the same across quantile levels.
- Estimate b_i and $\beta(\tau_k)$ for several quantiles simultaneously by solving

$$\min_{b,\beta} \sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_k \rho_\tau \left\{ y_{ij} - b_i - \mathbf{x}_{ij}^T \beta(\tau_k) \right\},$$

where w_k is the weight on the quantile level τ_k.

Penalized QR with Fixed Effects

- Since n is often large (relative to m_i), estimating a large number of b_i will inflate the variability of the estimates of other covariate effects.
- A fix: shrink b_i towards a common value through penalization.
- Minimize the following penalized objective function:

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_k \rho_\tau \{ y_{ij} - b_i - \mathbf{x}_{ij}^T \beta(\tau_k) \} + \lambda \sum_{i=1}^{n} |b_i|,$$

where λ is the penalization parameter:
  - $\lambda = 0$: no shrinkage
  - $\lambda = \infty$: $\hat{b}_i \to 0$.

Correlated Random Effects

- Abrevaya and Dahl (2008) adapt the Chamberlain (1982) correlated random effects model and estimate a model of birthweight, assuming that data contains exactly two births from a large number of mothers.
- Model:

$$y_{ij} = \mathbf{x}_{ij}^T \beta + b_i + e_{ij}, i = 1, \cdots, n, j = 1, 2, \quad (9)$$

where
  - b_i: unobservable mother effect;
  - $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$: covariate values from both births of a given mother;
  - b_i and $\mathbf{x}_{ij}$ are likely to be correlated.

Correlated Random Effects

- Assume the correlated-random-effect model (Chamberlain, 1982):

$$b_i = \theta + \mathbf{x}_{i1}^T \lambda_1 + \mathbf{x}_{i2}^T \lambda_2 + v_i,$$

where v_i is independent of $\mathbf{x}_{ij}$ and e_{ij}.
- Inserting the above expression to model (9) we get

$$\begin{aligned} y_{i1} &= \theta + \mathbf{x}_{i1}^T(\beta + \lambda_1) + \mathbf{x}_{i2}^T \lambda_2 + v_i + e_{ij}, \\ y_{i2} &= \theta + \mathbf{x}_{i1}^T \lambda_1 + \mathbf{x}_{i2}^T(\beta + \lambda_2) + v_i + e_{ij}, \end{aligned}$$

so

$$\begin{aligned} \beta &= \frac{\partial E(y_{i1}|\mathbf{x}_{i1}, \mathbf{x}_{i2})}{\partial \mathbf{x}_{i1}} - \frac{\partial E(y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2})}{\partial \mathbf{x}_{i1}} \\ &= \frac{\partial E(y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2})}{\partial \mathbf{x}_{i2}} - \frac{\partial E(y_{i1}|\mathbf{x}_{i1}, \mathbf{x}_{i2})}{\partial \mathbf{x}_{i2}}. \end{aligned}$$

Correlated Random Effects

- Analogous to the conditional mean model, assume the following models:

$$\begin{aligned} Q_\tau(y_{i1}|\mathbf{x}_{ij}) &= \theta_1(\tau) + \mathbf{x}_{i1}^T\{\beta(\tau) + \lambda_1(\tau)\} + \mathbf{x}_{i2}^T\lambda_2(\tau), \\ Q_\tau(y_{i2}|\mathbf{x}_{ij}) &= \theta_2(\tau) + \mathbf{x}_{i1}^T\lambda_1(\tau) + \mathbf{x}_{i2}^T\{\beta(\tau) + \lambda_2(\tau)\}. \end{aligned} \qquad (10)$$

- Can interpret $\beta(\tau)$ as

$$\begin{aligned} \beta(\tau) &= \frac{\partial Q_\tau(y_{i1}|\mathbf{x}_{i1},\mathbf{x}_{i2})}{\partial \mathbf{x}_{i1}} - \frac{\partial Q_\tau(y_{i2}|\mathbf{x}_{i1},\mathbf{x}_{i2})}{\partial \mathbf{x}_{i1}} \\ &= \frac{\partial Q_\tau(y_{i2}|\mathbf{x}_{i1},\mathbf{x}_{i2})}{\partial \mathbf{x}_{i2}} - \frac{\partial Q_\tau(y_{i1}|\mathbf{x}_{i1},\mathbf{x}_{i2})}{\partial \mathbf{x}_{i2}}. \end{aligned}$$

Correlated Random Effects

Parameters in (10) can be estimated with standard linear quantile regression.

The R package `rqpd` implements both this method and the penalized fixed effect approach. Available from R-Forge with the command:

```
install.packages("rqpd",repos="http://R-Forge.R-project.org")
```


Empirical Likelihood—Kim and Yang (2011)

- Model:

$$Y_{ij} = \mathbf{x}_{ij}^T\mathbf{b}_i(\tau) + e_{ij}, \qquad (11)$$

  where $\mathbf{b}_i(\tau) \sim N(\beta(\tau), \Sigma_i)$, $Q_{e_{ij}}(\tau|\mathbf{x}_{ij}) = 0$, and e_{ij} and $\mathbf{b}_i(\tau)$ are independent.
- Equivalent model:

$$Y_{ij} = \mathbf{x}_{ij}^T\beta(\tau) + \mathbf{x}_{ij}^T\mathbf{b}_i^*(\tau) + e_{ij}, \quad \mathbf{b}_i^*(\tau) \sim N(0, \Sigma_i). \qquad (12)$$

- $\beta(\tau)$: the average over cluster-specific quantile effects $\mathbf{b}_i(\tau)$.
- For notational simplicity, write $\beta(\tau) = \beta$, $\mathbf{b}_i(\tau) = \mathbf{b}_i$ and $\mathbf{b}_i^*(\tau) = \mathbf{b}_i^*$.

Cluster-specific quantile estimator of $\mathbf{b}_i$:

$$\widehat{\mathbf{b}}_i = \text{argmin}_{\mathbf{b}\in\mathbb{R}^q} \sum_{j=1}^{m_i} \rho_\tau(y_{ij} - \mathbf{x}_{ij}^T\mathbf{b}),$$

which is also a solution to the estimating equation

$$m_i^{-1/2} \sum_{j=1}^{m_i} \mathbf{x}_{ij}\psi_\tau(y_{ij} - \mathbf{x}_{ij}^T\mathbf{b}) \approx 0,$$

where $\psi_\tau(u) = \tau - I(u < 0)$.

Empirical likelihood (EL) for $\mathbf{b}_i(\tau)$ in the ith cluster:

$$\begin{aligned} L_{m_i}(\mathbf{b}_i) &= \max\Big\{\prod_{j=1}^{m_i} p_j \Big| \sum_{j=1}^{m_i} p_j\mathbf{x}_{ij}\psi_\tau(y_{ij} - \mathbf{x}_{ij}^T\mathbf{b}_i) = 0, \\ &\qquad \sum_{j=1}^{m_i} p_j = 1,\ 0 \leqslant p_j \leqslant 1\Big\}, \end{aligned}$$

where p_j is the weight for the jth observation in the ith cluster.

Kim and Yang (2011)

For a given $\mathbf{b}_i$, the EL $L_{m_i}(\mathbf{b}_i)$ can be calculated by using the standard Lagrange multiplier method; see Owen (2001), and Qin and Lawless (1994).

Notations

- $\boldsymbol{\sigma}$: collection of parameters involved in $\Sigma_1, \cdots, \Sigma_n$.
- $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma})$: collection of unknown parameters.
- $\boldsymbol{\theta}_0$: true value.
- $g_i(\mathbf{b}|\boldsymbol{\theta})$: the density of $\mathbf{b}_i$, that is, of $N(\boldsymbol{\beta}, \Sigma_i)$.

Kim and Yang (2011)

Semiparametric likelihood criterion function for $\mathbf{b}_i$

$$\widetilde{L}_{m_i}(\mathbf{b}|\boldsymbol{\theta}) = L_{m_i}(\mathbf{b})g_i(\mathbf{b}|\boldsymbol{\theta}).$$

Semiparametric likelihood criterion function for $\boldsymbol{\theta}$

$$\prod_{i=1}^{n} \int_{\mathbf{b}_i} \widetilde{L}_{m_i}(\mathbf{b}_i|\boldsymbol{\theta}) d\mathbf{b}_i.$$

Proposed estimators of $\boldsymbol{\theta}$ and $\mathbf{b}_i$:

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \int_{\mathbf{b}_i} \widetilde{L}_{m_i}(\mathbf{b}_i|\boldsymbol{\theta}) d\mathbf{b}_i,$$

$$\widetilde{\mathbf{b}}_i = \arg\max_{\mathbf{b}_i} \widetilde{L}_{m_i}(\mathbf{b}_i|\widetilde{\boldsymbol{\theta}}).$$

These estimators are computed by the quasi-posterior means using MCMC samplers.

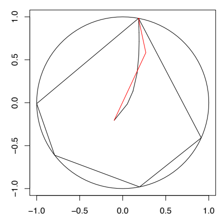
Kim and Yang (2011)

Properties

- As $m_i \to \infty$, the posterior mean estimator $\widetilde{\mathbf{b}}_i$ is a weighted average of $\widetilde{\beta}$ and the cluster-specific quantile regression estimator $\widehat{\mathbf{b}}_i$.
- As $m_i \to \infty$, the posterior mean estimator $\widetilde{\beta}$ is a weighted average of the cluster-specific QR estimator $\widehat{\mathbf{b}}_i$, and it is asymptotically normal.
- The variances of $\widetilde{\mathbf{b}}_i$ and $\widetilde{\theta}$ can be estimated by the variance-covariance matrices of the corresponding MCMC sequences. In contrast to conventional inference for quantile regression, the estimation of error density function is not required.
- Confidence intervals based on the normality or the percentiles of the MCMC sequence have approximately correct coverage as $m_i \to \infty$ for $\mathbf{b}_i$ and as $n \to \infty$ for $\boldsymbol{\theta}$. The result also holds when $\mathbf{b}_i$ have non-normal densities.

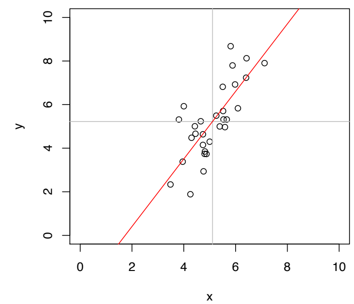
Appendix

Appendix 1
Quantile Regression Computation: From the Inside and the Outside



The Origin of Regression – Regression Through the Origin

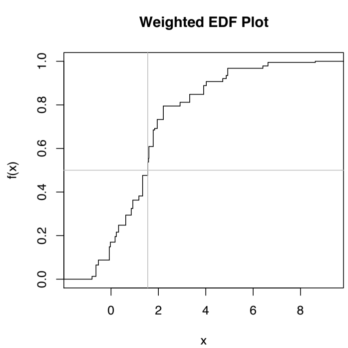
Find the line with mean residual zero that minimizes the sum of absolute residuals.



Problem: $\min_{\alpha,\beta} \sum_{i=1}^{n} |y_i - \alpha - x_i\beta|$ s.t. $\bar{y} = \alpha + \bar{x}\beta$.

Boscovich/Laplace *Methode de Situation*

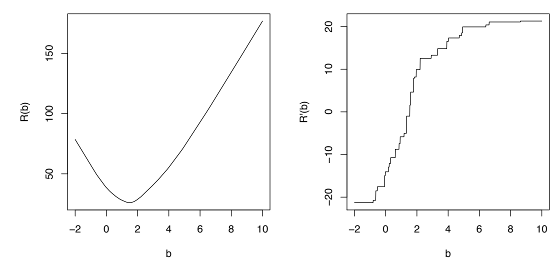
Algorithm: Order the n candidate slopes: $b_i = (y_i - \bar{y})/(x_i - \bar{x})$ denoting them by $b_{(i)}$ with associated weights $w_{(i)}$ where $w_i = |x_i - \bar{x}|$. Find the weighted median of these slopes.



Methode de Situation via Optimization

$$R(b) = \sum |\tilde{y}_i - \tilde{x}_i b| = \sum |\tilde{y}_i/\tilde{x}_i - b| \cdot |\tilde{x}_i|.$$

$$R'(b) = -\sum \text{sgn}(\tilde{y}_i/\tilde{x}_i - b) \cdot |\tilde{x}_i|.$$



Quantile Regression through the Origin in R

This can be easily generalized to compute quantile regression estimates:

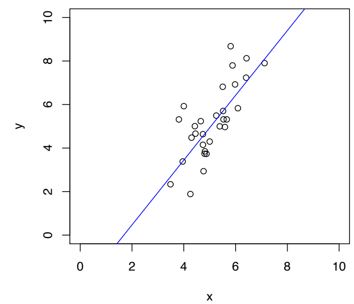
```
wquantile <- function(x, y, tau = 0.5) {
        o <- order(y/x)
        b <- (y/x)[o]
        w <- abs(x[o])
        k <- sum(cumsum(w) < ((tau - 0.5) * sum(x) + 0.5 * sum(w))
        list(coef = b[k + 1], k = ord[k+1])
}
```

Warning: When $\bar{x} = 0$ then τ is irrelevant. Why?

Edgeworth's (1888) Plural Median

What if we want to estimate both α and β by median regression?

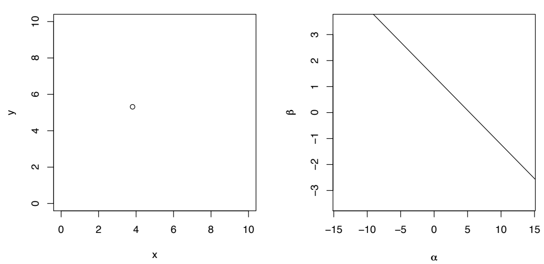
Problem: $\min_{\alpha,\beta} \sum_{i=1}^{n} |y_i - \alpha - x_i\beta|$



Edgeworth's (1888) Dual Plot: Anticipating Simplex

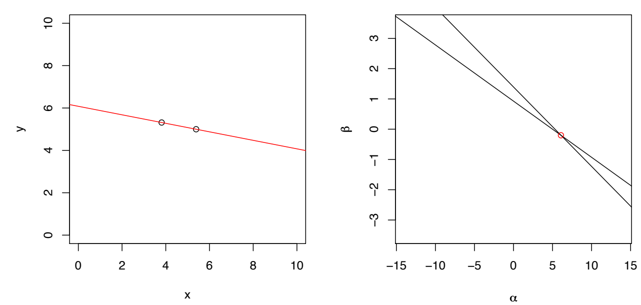
Points in sample space map to lines in parameter space.

$$(x_i, y_i) \mapsto \{(\alpha, \beta) : \alpha = y_i - x_i\beta\}$$



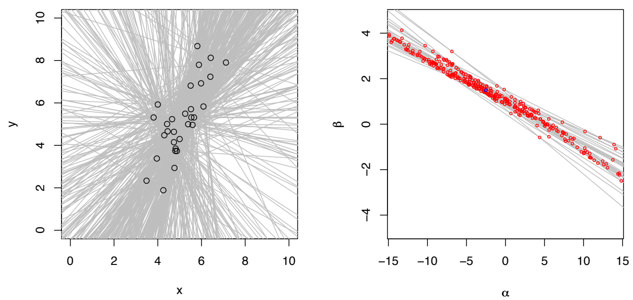
Edgeworth's (1888) Dual Plot: Anticipating Simplex

Lines through pairs of points in sample space map to points in parameter space.



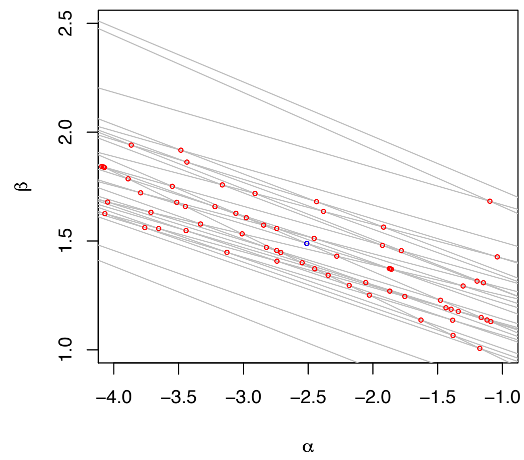
Edgeworth's (1888) Dual Plot: Anticipating Simplex

All pairs of observations produce $\binom{n}{2}$ points in dual plot.



Edgeworth's (1888) Dual Plot: Anticipating Simplex

Follow path of steepest descent through points in the dual plot.



Barrodale-Roberts Implementation of Edgeworth

```
rqx<- function(x, y, tau = 0.5, max.it = 50) { # Barrodale and Roberts -- lite
        p <- ncol(x); n <- nrow(x)
        h <- sample(1:n, size = p) #Phase I -- find a random (!) initial basis
        it <- 0
        repeat {
                it <- it + 1
                Xhinv <- solve(x[h,  ])
                bh <- Xhinv %*% y[h]
                rh <- y - x %*% bh
        #find direction of steepest descent along one of the edges
                g <-  - t(Xhinv) %*% t(x[ - h,  ]) %*% c(tau - (rh[ - h] < 0))
                g <- c(g + (1 - tau),  - g + tau)
                ming <- min(g)
                if(ming >= 0 || it > max.it) break
                h.out <- seq(along = g)[g == ming]
                sigma <- ifelse(h.out <= p, 1, -1)
                if(sigma < 0) h.out <- h.out - p
                d <- sigma * Xhinv[, h.out]
        #find step length by one-dimensional wquantile minimization
                xh <- x %*% d
                step <- wquantile(xh, rh, tau)
                h.in <- step$k
                h <- c(h[ - h.out], h.in)
        }
        if(it > max.it) warning("non-optimal solution: max.it exceeded")
        return(bh)
}
```


Linear Programming Duality

Primal: $\min_x\{c^\top x | Ax - b \in T,\ x \in S\}$
Dual: $\max_y\{b^\top y | c - A^\top y \in S^*,\ y \in T^*\}$

The sets S and T are closed convex cones, with dual cones S^* and T^*. A cone K^* is dual to K if:

$$K^* = \{y \in \mathbb{R}^n | x^\top y \geqslant 0 \ if\ x \in K\}$$

Note that for any feasible point (x, y)

$$b^\top y \leqslant y^\top Ax \leqslant c^\top x$$

while optimality implies that

$$b^\top y = c^\top x.$$

Quantile Regression Primal and Dual

Splitting the QR "residual" into positive and negative parts, yields the primal linear program,

$$\min_{(b,u,v)} \{\tau 1^\top u + (1-\tau)1^\top v \mid Xb + u - v - y \in \{0\}, \quad (b,u,v) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}\}.$$

with dual program:

$$\max_d \{y^\top d \mid X^\top d \in \{0\}, \quad \tau 1 - d \in \mathbb{R}_+^n, \quad (1-\tau)1 + d \in \mathbb{R}_+^n\},$$

$$\max_d \{y^\top d \mid X^\top d = 0, \; d \in [\tau - 1, \tau]^n\},$$

$$\max_a \{y^\top a \mid X^\top a = (1-\tau)X^\top 1, \quad a \in [0,1]^n\}$$

Quantile Regression Dual

The dual problem for quantile regression may be formulated as:

$$\max_a \{y^\top a | X^\top a = (1-\tau)X^\top 1, \; a \in [0,1]^n\}$$

What do these $\hat{a}_i(\tau)$'s mean statistically?
They are regression rank scores (Gutenbrunner and Jurečková (1992)):

$$\hat{a}_i(\tau) \in \begin{cases} \{1\} & \text{if} \quad y_i > x_i^\top \hat{\beta}(\tau) \\ (0,1) & \text{if} \quad y_i = x_i^\top \hat{\beta}(\tau) \\ \{0\} & \text{if} \quad y_i < x_i^\top \hat{\beta}(\tau) \end{cases}$$

The integral $\int \hat{a}_i(\tau) d\tau$ is something like the rank of the ith observation. It answers the question: On what quantile does the ith observation lie?

Linear Programming: The Inside Story

The Simplex Method (Edgeworth/Dantzig/Kantorovich) moves from vertex to vertex on the outside of the constraint set until it finds an optimum.
Interior point methods (Frisch/Karmarker/et al) take Newton type steps toward the optimal vertex from inside the constraint set.
A toy problem: Given a polygon inscribed in a circle, find the point on the polygon that maximizes the sum of its coordinates:

$$\max\{e^\top u | A^\top x = u, \; e^\top x = 1, \; x \geqslant 0\}$$

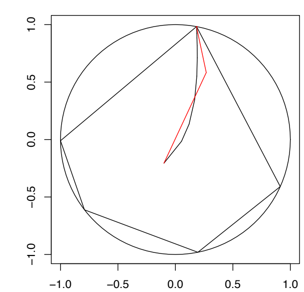
were e is vector of ones, and A has rows representing the n vertices.
Eliminating u, setting $c = Ae$, we can reformulate the problem as:

$$\max\{c^\top x | e^\top x = 1, \quad x \geqslant 0\},$$

Toy Story: From the Inside

Simplex goes around the outside of the polygon; interior point methods tunnel from the inside, solving a sequence of problems of the form:

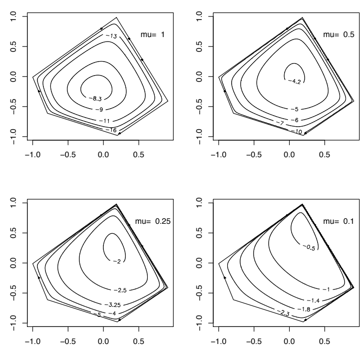
$$\max\{c^\top x + \mu \sum_{i=1}^n \log x_i | e^\top x = 1\}$$



Toy Story: From the Inside

By letting $\mu \to 0$ we get a sequence of smooth problems whose solutions approach the solution of the LP:

$$\max\{c^\top x + \mu \sum_{i=1}^{n} \log x_i | e^\top x = 1\}$$



Implementation: Meketon's Affine Scaling Algorithm

```
meketon <- function (x, y, eps = 1e-04, beta = 0.97) {
   f <- lm.fit(x,y)
   n <- length(y)
   w <- rep(0, n)
   d <- rep(1, n)
   its <- 0
   while(sum(abs(f$resid)) - crossprod(y, w) > eps) {
       its <- its + 1
       s <- f$resid * d
       alpha <- max(pmax(s/(1 - w), -s/(1 + w)))
       w <- w + (beta/alpha) * s
       d <- pmin(1 - w, 1 + w)^2
       f <- lm.wfit(x,y,d)
       }
   list(coef = f$coef, iterations = its)
   }
```


Mehrotra Primal-Dual Predictor-Corrector Algorithm

The algorithms implemented in `quantreg` for R are based on Mehrotra's Predictor-Corrector approach. Although somewhat more complicated than Meketon this has several advantages:

- Better numerical stability and efficiency due to better central path following,
- Easily generalized to incorporate linear inequality constraints.
- Easily generalized to exploit sparsity of the design matrix.

These features are all incorporated into various versions of the algorithm in `quantreg`, and coded in Fortran.

Back to Basics

Which is easier to compute: the median or the mean?

```
> x <- rnorm(100000000) # n = 10^8
> system.time(mean(x))
    user  system elapsed
  10.277   0.035  10.320
> system.time(kuantile(x,.5))
    user  system elapsed
   5.372   3.342   8.756
```

`kuantile` is a `quantreg` implementation of the Floyd-Rivest (1975) algorithm. For the median it requires $1.5n + O((n \log n)^{1/2})$ comparisons.

Portnoy and Koenker (1997) propose a similar strategy for "preprocessing" quantile regression problems to improve efficiency for large problems.

Globbing for Median Regression

Rather than solving $\min \sum |y_i - x_i b|$ consider:

1. Preliminary estimation using random $m = n^{2/3}$ subset,
2. Construct confidence band $x_i^\top \hat{\beta} \pm \kappa \| \hat{V}^{1/2} x_i \|$.
3. Find $J_L = \{i | y_i \text{ below band }\}$, and $J_H = \{i | y_i \text{ above band }\}$,
4. Glob observations together to form pseudo observations:

$$(x_L, y_L) = (\sum_{i \in J_L} x_i, -\infty), \quad (x_H, y_H) = (\sum_{i \in J_H} x_i, +\infty)$$

5. Solve the problem (with m+2 observations)

$$\min \sum |y_i - x_i b| + |y_L - x_L b| + |y_H - x_H b|$$

6. Verify that globbed observations have the correct predicted signs.

The Laplacian Tortoise and the Gaussian Hare

Retouched 18th century woodblock photo-print

Appendix 2
Quantile Autoregression

Based on joint work with Zhijie Xiao, Boston College.

Outline

- A Motivating Example
- The QAR Model
- Inference for QAR models
- Forecasting with QAR Models
- Surgeon General's Warning
- Conclusions

Introduction

In classical regression and autoregression models

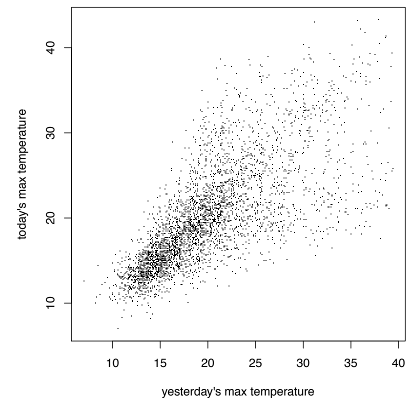
$$y_i = h(x_i, \theta) + u_i,$$
$$y_t = \alpha y_{t-1} + u_t$$

conditioning covariates influence only the location of the conditional distribution of the response:

Response $=$ Signal $+$ IID Noise.

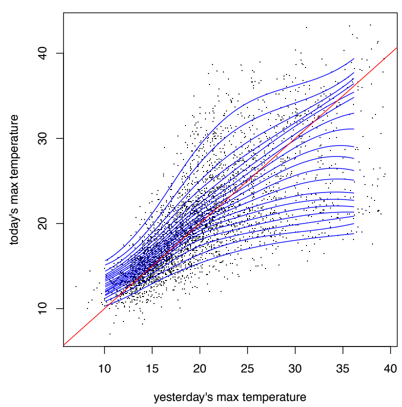
But why should noise always be so well-behaved?

A Motivating Example



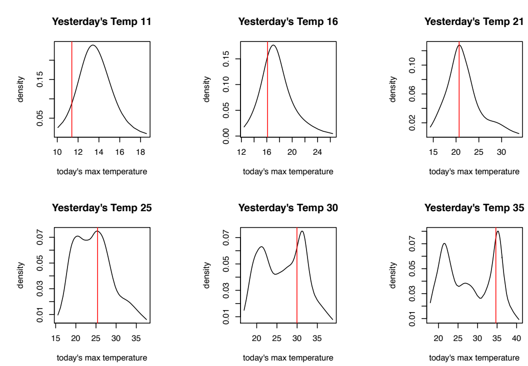
Daily Temperature in Melbourne: An AR(1) Scatterplot

Estimated Conditional Quantiles of Daily Temperature



Daily Temperature in Melbourne: A Nonlinear QAR(1) Model

Conditional Densities of Melbourne Daily Temperature



Location, scale and shape all change with y_{t-1}.
When today is hot, tomorrow's temperature is bimodal!

Linear AR(1) and QAR(1) Models

The classical linear AR(1) model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t,$$

with iid errors, $u_t : t = 1, \cdots, T$, implies

$$E(y_t | \mathcal{F}_{t-1}) = \alpha_0 + \alpha_1 y_{t-1}$$

and conditional quantile functions are all parallel:

$$Q_{y_t}(\tau | \mathcal{F}_{t-1}) = \alpha_0(\tau) + \alpha_1 y_{t-1}$$

with $\alpha_0(\tau) = F_u^{-1}(\tau)$ just the quantile function of the u_t's.
But isn't this rather boring? What if we let α_1 depend on τ too?

A Random Coefficient Interpretation

If the conditional quantiles of the response satisfy:

$$Q_{y_t}(\tau | \mathcal{F}_{t-1}) = \alpha_0(\tau) + \alpha_1(\tau) y_{t-1}$$

then we can generate responses from the model by replacing τ by uniform random variables:

$$y_t = \alpha_0(u_t) + \alpha_1(u_t) y_{t-1} \quad u_t \sim \text{iid } U[0,1].$$

This is a very special form of random coefficient autoregressive (RCAR) model with comonotonic coefficients.

On Comonotonicity

Definition: Two random variables $X, Y : \Omega \to \mathbb{R}$ are comonotonic if there exists a third random variable $Z : \Omega \to \mathbb{R}$ and increasing functions f and g such that $X = f(Z)$ and $Y = g(Z)$.

- If X and Y are comonotonic they have rank correlation one.
- From our point of view the crucial property of comonotonic random variables is the behavior of quantile functions of their sums, X, Y comonotonic implies:

$$F_{X+Y}^{-1}(\tau) = F_X^{-1}(\tau) + F_Y^{-1}(\tau)$$

- X and Y are driven by the same random (uniform) variable.

The QAR(p) Model

Consider a p-th order QAR process,

$$Q_{y_t}(\tau | \mathcal{F}_{t-1}) = \alpha_0(\tau) + \alpha_1(\tau) y_{t-1} + \ldots + \alpha_p(\tau) y_{t-p}$$

Equivalently, we have random coefficient model,

$$\begin{aligned} y_t &= \alpha_0(u_t) + \alpha_1(u_t) y_{t-1} + \cdots + \alpha_p(u_t) y_{t-p} \\ &\equiv x_t^\top \alpha(u_t). \end{aligned}$$

Now, all $p+1$ random coefficients are comonotonic, functionally dependent on the same uniform random variable.

Vector QAR(1) representation of the QAR(p) Model

$$Y_t = \mu + A_t Y_{t-1} + V_t$$

where

$$\mu = \begin{bmatrix} \mu_0 \\ 0_{p-1} \end{bmatrix}, A_t = \begin{bmatrix} a_t & \alpha_p(u_t) \\ I_{p-1} & 0_{p-1} \end{bmatrix}, V_t = \begin{bmatrix} v_t \\ 0_{p-1} \end{bmatrix}$$

$$a_t = [\alpha_1(u_t), \dots, \alpha_{p-1}(u_t)],$$

$$Y_t = [y_t, \cdots, y_{t-p+1}]^\top,$$

$$v_t = \alpha_0(u_t) - \mu_0.$$

It all looks rather complex and multivariate, but it is really still nicely univariate and very tractable.

Slouching Toward Asymptopia

We maintain the following regularity conditions:

- A.1 $\{v_t\}$ are iid with mean 0 and variance $\sigma^2 < \infty$. The CDF of v_t, F, has a continuous density f with $f(v) > 0$ on $\mathcal{V} = \{v : 0 < F(v) < 1\}$.
- A.2 Eigenvalues of $\Omega_A = E(A_t \otimes A_t)$ have moduli less than unity.
- A.3 Denote the conditional CDF $\Pr[y_t < y|\mathcal{F}_{t-1}]$ as $F_{t-1}(y)$ and its derivative as $f_{t-1}(y)$, f_{t-1} is uniformly integrable on $\mathcal{V}$.

Stationarity

Theorem 1: Under assumptions A.1 and A.2, the QAR(p) process y_t is covariance stationary and satisfies a central limit theorem

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (y_t - \mu_y) \Rightarrow N\left(0, \omega_y^2\right),$$

with

$$\begin{aligned} \mu_y &= \frac{\mu_0}{1 - \sum_{j=1}^{p} \mu_p}, \\ \mu_j &= E(\alpha_j(u_t)), \quad j = 0, ..., p, \\ \omega_y^2 &= \lim \frac{1}{n} E[\sum_{t=1}^{n} (y_t - \mu_y)]^2. \end{aligned}$$

Example: The QAR(1) Model

For the QAR(1) model,

$$Q_{y_t}(\tau|y_{t-1}) = \alpha_0(\tau) + \alpha_1(\tau) y_{t-1},$$

or with u_t iid $U[0,1]$.

$$y_t = \alpha_0(u_t) + \alpha_1(u_t) y_{t-1},$$

if $\omega^2 = E(\alpha_1^2(u_t)) < 1$, then y_t is covariance stationary and

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (y_t - \mu_y) \Rightarrow N\left(0, \omega_y^2\right),$$

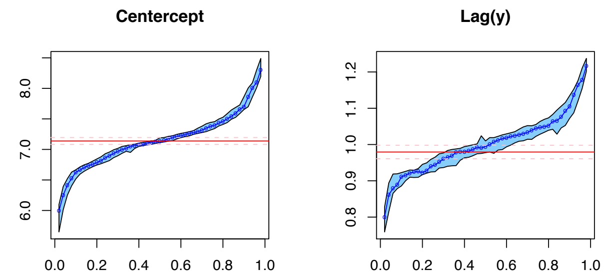
where $\mu_0 = E\alpha_0(u_t)$, $\mu_1 = E(\alpha_1(u_t))$, $\sigma^2 = V(\alpha_0(u_t))$, and

$$\mu_y = \frac{\mu_0}{(1 - \mu_1)}, \quad \omega_y^2 = \frac{(1 + \mu_1)\sigma^2}{(1 - \mu_1)(1 - \omega^2)},$$

Qualitative Behavior of QAR(p) Processes

- The model can exhibit unit-root-like tendencies, even temporarily explosive behavior, but episodes of mean reversion are sufficient to insure stationarity.
- Under certain conditions,the QAR(p) process is a semi-strong ARCH(p) process in the sense of Drost and Nijman (1993).
- The impulse response of y_{t+s} to a shock u_t is stochastic but converges (to zero) in mean square as $s \to \infty$.

Estimated QAR(1) v. AR(1) Models of U.S. Interest Rates



Data: Seasonally adjusted monthly: April, 1971 to June, 2002.
Do 3-month T-bills really have a unit root?

Estimation of the QAR model

Estimation of the QAR models involves solving,

$$\hat{\alpha}(\tau) = \text{argmin}_\alpha \sum_{t=1}^{n} \rho_\tau(y_t - x_t^\top \alpha),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, the $\surd$-function.
Fitted conditional quantile functions of y_t, are given by,

$$\hat{Q}_t(\tau|x_t) = x_t^\top \hat{\alpha}(\tau),$$

and conditional densities by the difference quotients,

$$\hat{f}_t(\tau|x_{t-1}) = \frac{2h}{\hat{Q}_t(\tau + h|x_{t-1}) - \hat{Q}_t(\tau - h|x_{t-1})},$$

The QAR Process

Theorem 2: Under our regularity conditions,

$$\sqrt{n}\Omega^{-1/2}(\hat{\alpha}(\tau) - \alpha(\tau)) \Rightarrow B_{p+1}(\tau),$$

a $(p+1)$-dimensional standard Brownian Bridge, with

$$\begin{aligned}
\Omega &= \Omega_1^{-1}\Omega_0\Omega_1^{-1}. \\
\Omega_0 &= E(x_t x_t^\top) = \lim n^{-1} \sum_{t=1}^{n} x_t x_t^\top, \\
\Omega_1 &= \lim n^{-1} \sum_{t=1}^{n} f_{t-1}(F_{t-1}^{-1}(\tau)) x_t x_t^\top.
\end{aligned}$$

Inference for QAR models

For fixed $\tau = \tau_0$ we can test the hypothesis:

$$H_0: \quad R\alpha(\tau) = r$$

using the Wald statistic,

$$W_n(\tau) = \frac{n(R\hat{\alpha}(\tau) - r)^\top [R\hat{\Omega}_1^{-1}\hat{\Omega}_0\hat{\Omega}_1^{-1}R^\top]^{-1}(R\hat{\alpha}(\tau) - r)}{\tau(1-\tau)}$$

This approach can be extended to testing on general index sets $\tau \in \mathcal{T}$ with the corresponding Wald process.

Asymptotic Inference

Theorem: Under H_0, $W_n(\tau) \Rightarrow Q_m^2(\tau)$, where $Q_m(\tau)$ is a Bessel process of order $m = \text{rank}(R)$. For fixed τ, $Q_m^2(\tau) \sim \chi_m^2$.

- Kolmogorov-Smirov or Cramer-von-Mises statistics based on $W_n(\tau)$ can be used to implement the tests.
- For known R and r this leads to a very nice theory – estimated R and/or r testing raises new questions.
- The situation is quite analogous to goodness-of-fit testing with estimated parameters.

Example: Unit Root Testing

Consider the augmented Dickey-Fuller model

$$y_t = \delta_0 + \delta_1 y_{t-1} + \sum_{j=2}^{p} \delta_j \Delta y_{t-j} + u_t.$$

We would like to test this constant coefficients version of the model against the more general QAR(p) version:

$$Q_{y_t}(\tau|x_t) = \delta_0(\tau) + \delta_1(\tau)y_{t-1} + \sum_{j=2}^{p} \delta_j(\tau)\Delta y_{t-j}$$

The hypothesis: $H_0 : \delta_1(\tau) = \bar{\delta}_1 = 1$, for $\tau \in \mathcal{T} = [\tau_0, 1-\tau_0]$, is considered in Koenker and Xiao (JASA, 2004).

Example: Two Tests

- When $\bar{\delta}_1 < 1$ is known we have the candidate process,

  $$V_n(\tau) = \sqrt{n}(\hat{\delta}_1(\tau) - \bar{\delta}_1)/\hat{\omega}_{11}.$$

  where $\hat{\omega}_{11}^2$ is the appropriate element from $\hat{\Omega}_1^{-1}\hat{\Omega}_0\hat{\Omega}_1^{-1}$. Fluctuations in $V_n(\tau)$ can be evaluated with the Kolmogorov-Smirnov statistic,

  $$\sup_{\tau \in \mathcal{T}} V_n(\tau) \Rightarrow \sup_{\tau \in \mathcal{T}} B(\tau).$$

- When $\bar{\delta}_1$ is unknown we may replace it with an estimate, but this disrupts the convenient asymptotic behavior. Now,

  $$\hat{V}_n(\tau) = \sqrt{n}((\hat{\delta}_1(\tau) - \bar{\delta}_1) - (\hat{\delta}_1 - \bar{\delta}_1))/\hat{\omega}_{11}$$

Martingale Transformation of $\hat{V}_n(\tau)$

Khmaladze (1981) suggested a general approach to the transformation of parametric empirical processes like $\hat{V}_n(\tau)$:

$$\widetilde{V}_n(\tau) = \hat{V}_n(\tau) - \int_0^\tau \left[\dot{g}_n(s)^\top C_n^{-1}(s) \int_s^1 \dot{g}_n(r) d\hat{V}_n(r) \right] ds$$

where $\dot{g}_n(s)$ and $C_n(s)$ are estimators of

$$\dot{g}(r) = (1, (\dot{f}/f)(F^{-1}(r)))^\top; \, C(s) = \int_s^1 \dot{g}(r)\dot{g}(r)^\top dr.$$

This is a generalization of the classical Doob-Meyer decomposition.

Restoration of the ADF property

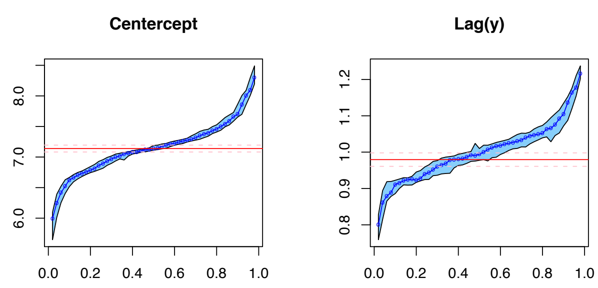
Theorem Under H_0, $\tilde{V}_n(\tau) \Rightarrow W(\tau)$ and therefore

$$\sup_{\tau \in \mathcal{T}} \|\tilde{V}_n(\tau)\| \Rightarrow \sup_{\tau \in \mathcal{T}} \|W(\tau)\|,$$

with $W(r)$ a standard Brownian motion.

- The martingale transformation of Khmaladze annihilates the contribution of the estimated parameters to the asymptotic behavior of the $\hat{V}_n(\tau)$ process, thereby restoring the asymptotically distribution free (ADF) character of the test.

Three Month T-Bills Again



A test of the "location-shift" hypothesis yields a test statistic of 2.76 which has a p-value of roughly 0.01, contradicting the conclusion of the conventional Dickey-Fuller test.

QAR Models for Longitudinal Data

- In estimating growth curves it is often valuable 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.
- Finnish Height Data: $\{Y_i(t_{i,j}) : \; j = 1, \ldots, J_i, \; i = 1, \ldots, n.\}$
- Partially Linear Model [Pere, Wei, Koenker, and He (2006)]:

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

Parametric Components of the 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)

Estimates of the QAR(1) parameters, $\alpha(\tau)$ and $\beta(\tau)$ and the mid-parental height effect, $\gamma(\tau)$, for Finnish children ages 0 to 2 years.

Forecasting with QAR Models

Given an estimated QAR model,

$$\hat{Q}_{y_t}(\tau|\mathcal{F}_{t-1}) = x_t^\top \hat{\alpha}(\tau)$$

based on data: $y_t : t = 1, 2, \cdots, T$, we can forecast

$$\hat{y}_{T+s} = \tilde{x}_{T+s}^\top \hat{\alpha}(U_s), \ s = 1, \cdots, S,$$

where $\tilde{x}_{T+s} = [1, \tilde{y}_{T+s-1}, \cdots, \tilde{y}_{T+s-p}]^\top$, $U_s \sim U[0, 1]$, and

$$\tilde{y}_t = \begin{cases} y_t & \text{if} \quad t \leqslant T, \\ \hat{y}_t & \text{if} \quad t > T. \end{cases}$$

Conditional density forecasts can be made based on an ensemble of such forecast paths.

Linear QAR Models May Pose Statistical Health Risks

- Lines with distinct slopes eventually intersect. [Euclid: P5]
- Quantile functions, $Q_Y(\tau|x)$ should be monotone in τ for all x, intersections imply point masses – or even worse.
- What is to be done?
  - Constrained QAR: Quantiles can be estimated simultaneously subject to linear inequality restrictions.
  - Nonlinear QAR: Abandon linearity in the lagged y_t's, as in the Melbourne temperature example, both parametric and nonparametric options are available.

Nonlinear QAR Models via Copulas

An interesting class of stationary, Markovian models can be expressed in terms of their copula functions:

$$G(y_t, y_{t-1}, \cdots, y_{y-p}) = C(F(y_t), F(y_{t-1}), \cdots, F(y_{y-p}))$$

where G is the joint df and F the common marginal df.

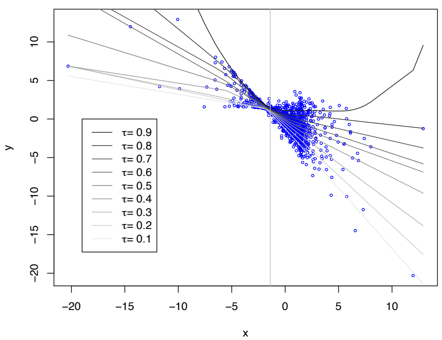
- Differentiating, $C(u, v)$, with respect to u, gives the conditional df,

$$H(y_t|y_{t-1}) = \frac{\partial}{\partial u} C(u, v)|_{(u=F(y_t), v=F(y_{t-1}))}$$

- Inverting we have the conditional quantile functions,

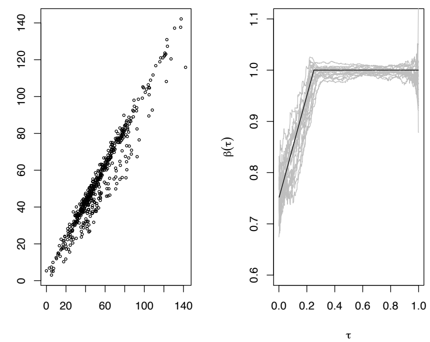
$$Q_{y_t}(\tau|y_{t-1}) = h(y_{t-1}, \theta(\tau))$$

Example 1 (Fan and Fan)



Model: $Q_{y_t}(\tau|y_{t-1}) = -(1.7 - 1.8\tau)y_{t-1} + \Phi^{-1}(\tau)$.

Example 2 (Near Unit Root)



Model: $Q_{y_t}(\tau|y_{t-1}) = 2 + \min\{\frac{3}{4} + \tau, 1\}y_{t-1} + 3\Phi^{-1}(\tau)$.

Conclusions

- QAR models are an attempt to expand the scope of classical linear time-series models permitting lagged covariates to influence scale and shape as well as location of conditional densities.
- Efficient estimation via familiar linear programming methods.
- Random coefficient interpretation nests many conventional models including ARCH.
- Wald-type inference is feasible for a large class of hypotheses; rank based inference is also an attractive option.
- Forecasting conditional densities is potentially valuable.
- Many new and challenging open problems. . . .

Appendix 3
Endogoneity and All That

Is there IV for QR?

- Amemiya (1982) and Powell (1983) consider analogues of 2SLS for median regression models
- Chen and Portnoy (1986) consider extensions to quantile regression
- Abadie, Angrist and Imbens (2002) consider models with binary endogonous treatment
- Chernozhukov and Hansen (2003) propose "inverse" quantile regression
- Chesher (2003) considers triangular models with continuous endogonous variables.

Chernozhukov and Hansen QRIV

Motivation: Yet another way to view two stage least squares.

Model: $y = X\beta + Z\alpha + u, \quad W \perp\!\!\!\perp u$

Estimator:

$$\hat{\alpha} = \text{argmin}_{\alpha} \|\hat{\gamma}(\alpha)\|^2_{A=W^\top M_X W}$$

$$\hat{\gamma}(\alpha) = \text{argmin}_{\gamma} \|y - X\beta - Z\alpha - W\gamma\|^2$$

Thm $\hat{\alpha} = (Z^\top P_{M_X W} Z)^{-1} Z^\top P_{M_X W} y$, the 2SLS estimator.

Heuristic: $\hat{\alpha}$ is chosen to make $\|\hat{\gamma}(\alpha)\|$ as small as possible to satisfy (approximately) the exclusion restriction/assumption.

Generalization: The quantile regression version simply replaces $\|\cdot\|^2$ in the definition of $\hat{\gamma}$ by the corresponding QR norm.

A Linear Location Shift Recursive Model

$$Y = S\alpha_1 + x^\top \alpha_2 + \epsilon + \lambda\nu \quad (13)$$

$$S = z\beta_1 + x^\top \beta_2 + \nu \quad (14)$$

Suppose: $\epsilon \perp\!\!\!\perp \nu$ and $(\epsilon, \nu) \perp\!\!\!\perp (z, x)$. Substituting for ν from (2) into (1),

$$\begin{aligned} Q_Y(\tau_1|S,x,z) &= S(\alpha_1 + \lambda) + x^\top(\alpha_2 - \lambda\beta_2) + z(-\lambda\beta_1) + F_\epsilon^{-1}(\tau_1) \\ Q_S(\tau_2|z,x) &= z\beta_1 + x^\top\beta_2 + F_\nu^{-1}(\tau_2) \end{aligned}$$

$$\begin{aligned} \pi_1(\tau_1, \tau_2) &= \nabla_{S_i} Q_{Y_i}|_{S_i = Q_{S_i}} + \frac{\nabla_{z_i} Q_{Y_i}|_{S_i = Q_{S_i}}}{\nabla_{z_i} Q_{S_i}} \\ &= (\alpha_1 + \lambda) + (-\lambda\beta_1)/\beta_1 \\ &= \alpha_1 \end{aligned}$$

A Linear Location-Scale Shift Model

$$\begin{aligned} Y &= S\alpha_1 + x^\top\alpha_2 + S(\epsilon + \lambda\nu) \\ S &= z\beta_1 + x^\top\beta_2 + \nu \\ \pi_1(\tau_1, \tau_2) &= \alpha_1 + F_\epsilon^{-1}(\tau_1) + \lambda F_\nu^{-1}(\tau_2) \end{aligned}$$

$$\begin{aligned} Q_Y(\tau_1|S,x,z) &= S\theta_1(\tau_1) + x^\top\theta_2 + S^2\theta_3 + Sz\theta_4 + Sx^\top\theta_5 \\ Q_S(\tau_2|z,x) &= z\beta_1 + x^\top\beta_2 + F_\nu^{-1}(\tau_2) \end{aligned}$$

$$\hat{\pi}_1(\tau_1,\tau_2) = \sum_{i=1}^n w_i \Big\{ \hat{\theta}_1(\tau_1) + 2\hat{Q}_{S_i}\hat{\theta}_3(\tau_1) + z_i\hat{\theta}_4(\tau_1) + x_i^\top\hat{\theta}_5(\tau_1) + \frac{\hat{Q}_{S_i}\hat{\theta}_4(\tau_1)}{\hat{\beta}_1(\tau_2)} \Big\}$$

a weighted average derivative estimator with $\hat{Q}_{S_i} = \hat{Q}_S(\tau_2|z_i, x_i)$.

The General Recursive Model

$$
\begin{aligned}
Y &= \varphi_1(S, x, \epsilon, \nu;\ \alpha) \\
S &= \varphi_2(z, x, \nu;\ \beta)
\end{aligned}
$$

Suppose: $\epsilon \perp\!\!\!\perp \nu$ and $(\epsilon, \nu) \perp\!\!\!\perp (z, x)$. Solving for ν and substituting we have the conditional quantile functions,

$$
\begin{aligned}
Q_Y(\tau_1 | S, x, z) &= h_1(S, x, z, \theta(\tau_1)) \\
Q_S(\tau_2 | z, x) &= h_2(z, x, \beta(\tau_2))
\end{aligned}
$$

Extensions to more than two endogonous variables are "straightforward."

The (Chesher) Weighted Average Derivative Estimator

$$
\begin{aligned}
\hat\theta(\tau_1) &= \text{argmin}_\theta \sum_{i=1}^n \rho_{\tau_1}(Y_i - h_1(S, x, z, \theta(\tau_1))) \\
\hat\beta(\tau_2) &= \text{argmin}_\beta \sum_{i=1}^n \rho_{\tau_2}(S_i - h_2(z, x, \beta(\tau_2)))
\end{aligned}
$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, giving structural estimators:

$$
\hat\pi_1(\tau_1, \tau_2) = \sum_{i=1}^n w_i \left\{ \nabla_S \hat h_{1i}|_{S_i = \hat h_{2i}} + \frac{\nabla_z \hat h_{1i}|_{S_i = \hat h_{2i}}}{\nabla_z \hat h_{2i}} \right\},
$$

$$
\hat\pi_2(\tau_1, \tau_2) = \sum_{i=1}^n w_i \left\{ \nabla_x \hat h_{1i}|_{S_i = \hat h_{2i}} - \frac{\nabla_z \hat h_{1i}|_{S_i = \hat h_{2i}}}{\nabla_z \hat h_{2i}} \nabla_x \hat h_{2i} \right\},
$$

2SLS as a Control Variate Estimator

$$
\begin{aligned}
Y &= S\alpha_1 + X_1\alpha_2 + u \equiv Z\alpha + u \\
S &= X\beta + V, \quad \text{where } X = [X_1 \vdots X_2]
\end{aligned}
$$

Set $\hat V = S - \hat S \equiv M_X Y_1$, and consider the least squares estimator of the model,

$$
Y = Z\alpha + \hat V \gamma + w
$$

Claim: $\hat\alpha_{CV} \equiv (Z^\top M_{\hat V} Z)^{-1} Z^\top M_{\hat V} Y = (Z^\top P_X Z)^{-1} Z^\top P_X Y \equiv \hat\alpha_{2SLS}$.

Proof of Control Variate Equivalence

$$
M_{\hat V} = M_{M_X S} = I - M_X S (S^\top M_X S)^{-1} S^\top M_X
$$

$$
\begin{aligned}
S^\top M_{\hat V} &= S^\top - S^\top M_X = S^\top P_X \\
X_1^\top M_{\hat V} &= X_1^\top - X_1^\top M_X = X_1^\top = X_1^\top P_X
\end{aligned}
$$

Reward for information leading to a reference prior to Dhrymes (1970). Recent work on the control variate approach by Blundell, Powell, Smith, Newey and others.

Quantile Regression Control Variate Estimation I

Location scale shift model:

$$\begin{aligned} Y &= S(\alpha_1 + \epsilon + \lambda\nu) + x^\top \alpha_2 \\ S &= z\beta_1 + x^\top \beta_2 + \nu. \end{aligned}$$

Using $\hat{\nu}(\tau_2) = S - \hat{Q}_S(\tau_2|z,x)$ as a control variate,

$$\begin{aligned} Y &= w^\top \alpha(\tau_1,\tau_2) + \lambda S(\hat{Q}_S - Q_S) + S(\epsilon - F_\epsilon^{-1}(\tau_1)), \\ \text{where} \quad & w^\top = (S, x^\top, S\hat{\nu}(\tau_2)) \\ & \alpha(\tau_1,\tau_2) = (\alpha_1(\tau_1,\tau_2), \alpha_2, \lambda)^\top \\ & \alpha_1(\tau_1,\tau_2) = \alpha_1 + F_\epsilon^{-1}(\tau_1) + \lambda F_\nu^{-1}(\tau_2). \end{aligned}$$

$$\hat{\alpha}(\tau_1,\tau_2) = \text{argmin}_a \sum_{i=1}^n \rho_{\tau_1}(Y_i - w_i^\top a).$$

Quantile Regression Control Variate Estimation II

$$\begin{aligned} Y &= \varphi_1(S, x, \epsilon, \nu;\ \alpha) \\ S &= \varphi_2(z, x, \nu;\ \beta) \end{aligned}$$

Regarding $\nu(\tau_2) = \nu - F_\nu^{-1}(\tau_2)$ as a control variate, we have

$$\begin{aligned} Q_Y(\tau_1|S, x, \nu(\tau_2)) &= g_1(S, x, \nu(\tau_2), \alpha(\tau_1,\tau_2)) \\ Q_S(\tau_2|z, x) &= g_2(z, x, \beta(\tau_2)) \end{aligned}$$

$$\hat{\nu}(\tau_2) = \varphi_2^{-1}(S, z, x, \hat{\beta}) - \varphi_2^{-1}(\hat{Q}_s, z, x, \hat{\beta})$$

$$\hat{\alpha}(\tau_1,\tau_2) = \text{argmin}_a \sum_{i=1}^n \rho_{\tau_1}(Y_i - g_1(S, x, \hat{\nu}(\tau_2), a)).$$

Asymptopia

Theorem: Under regularity conditions, the weighted average derivative and control variate estimators of the Chesher structural effect have an asymptotic linear (Bahadur) representation, and after efficient reweighting of both estimators, the control variate estimator has smaller covariance matrix than the weighted average derivative estimator.

Remark: The control variate estimator imposes more stringent restrictions on the estimation of the hybrid structural equation and should thus be expected to perform better when the specification is correct. The advantages of the control variate approach are magnified in situations of overidentification.

Asymptotics for WAD

Theorem

The $\hat{\pi}_n(\tau_1,\tau_2)$ has the asymptotic linear (Bahadur) representation,

$$\begin{aligned} \sqrt{n}(\hat{\pi}_n(\tau_1,\tau_2) - \pi(\tau_1,\tau_2)) = \ & W_1 \bar{J}_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma_{i1} \dot{h}_{i1} \psi_{\tau_1}(Y_{i1} - \xi_{i1}) \\ & + W_2 \bar{J}_2^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma_{i2} \dot{h}_{i2} \psi_{\tau_2}(Y_{i2} - \xi_{i2}) \\ & \Longrightarrow \mathcal{N}(0,\ \omega_{11} W_1 \bar{J}_1^{-1} J_1 \bar{J}_1^{-1} W_1^\top + \omega_{22} W_2 \bar{J}_2^{-1} J_2 \bar{J}_2^{-1} W_2^\top) \end{aligned}$$

$$\begin{aligned} & J_j = \lim_{n\to\infty} \frac{1}{n} \sum \sigma_{ij}^2 \dot{h}_{ij} \dot{h}_{ij}^\top, \quad \bar{J}_j = \lim_{n\to\infty} \frac{1}{n} \sum \sigma_{ij} f_{ij}(\xi_{ij}) \dot{h}_{ij} \dot{h}_{ij}^\top, \\ & W_1 = \nabla_\theta \pi(\tau_1,\tau_2), \quad W_2 = \nabla_\beta \pi(\tau_1,\tau_2), \\ & \dot{h}_{i1} = \nabla_\theta h_{i1}, \quad \dot{h}_{i2} = \nabla_\beta h_{i2}, \quad \omega_{jj} = \tau_j(1-\tau_j). \end{aligned}$$

Asymptotics for CV

Theorem

The $\hat{\alpha}_n(\tau_1, \tau_2)$ has the Bahadur representation,

$$\sqrt{n}(\hat{\alpha}_n(\tau_1,\tau_2) - \alpha(\tau_1,\tau_2)) = \bar{D}_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_{i1}\dot{g}_{i1}\psi_{\tau_1}(Y_{i1} - \xi_{i1})$$
$$+ \bar{D}_1^{-1}\bar{D}_{12}\bar{D}_2^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_{i2}\dot{g}_{i2}\psi_{\tau_2}(Y_{i2} - \xi_{i2})$$
$$\Longrightarrow \mathcal{N}(0,\ \omega_{11}\bar{D}_1^{-1}D_1\bar{D}_1^{-1} + \omega_{22}\bar{D}_1^{-1}\bar{D}_{12}\bar{D}_2^{-1}D_2\bar{D}_2^{-1}\bar{D}_{12}^\top\bar{D}_1^{-1})$$

$$D_j = \lim_{n\to\infty} n^{-1}\sum \sigma_{ij}^2\dot{g}_{ij}\dot{g}_{ij}^\top,\ \bar{D}_j = \lim_{n\to\infty} n^{-1}\sum \sigma_{ij}f_{ij}(\xi_{ij})\dot{g}_{ij}\dot{g}_{ij}^\top,$$
$$\bar{D}_{12} = \lim_{n\to\infty} n^{-1}\sum \sigma_{i1}f_{i1}\eta_i\dot{g}_{i1}\dot{g}_{i2}^\top,$$
$$\dot{g}_{i1} = \nabla_\alpha g_{i1},\ \dot{g}_{i2} = \nabla_\beta g_{i2},\ \eta_i = (\partial g_{1i}/\partial\nu_{i2}(\tau_2))(\nabla_{\nu_{i2}}\varphi_{i2})^{-1}.$$

ARE of WAD and CV

- Efficient weights: $\sigma_{ij} = f_{ij}(\xi_{ij})$

$$\sqrt{n}(\hat{\pi}_n(\tau_1,\tau_2) - \pi(\tau_1,\tau_2)) \Rightarrow \mathcal{N}(0, \omega_{11}W_1J_1^{-1}W_1^\top + \omega_{22}W_2J_2^{-1}W_2^\top)$$
$$\sqrt{n}(\hat{\alpha}_n(\tau_1,\tau_2) - \alpha(\tau_1,\tau_2)) \Rightarrow \mathcal{N}(0, \omega_{11}D_1^{-1} + \omega_{22}D_1^{-1}D_{12}D_2^{-1}D_{12}^\top D_1^{-1}).$$

The mapping: $\tilde{\pi}_n = L\hat{\alpha}_n,\ L\alpha = \pi.$

$$W_1J_1^{-1}W_1^\top \geqslant LD_1^{-1}L^\top$$
$$W_2J_2^{-1}W_2^\top \geqslant LD_1^{-1}D_{12}D_2^{-1}D_{12}^\top D_1^{-1}L^\top.$$

Theorem

Under efficient reweighting of both estimators,

$$Avar(\sqrt{n}\tilde{\pi}_n) \leqslant Avar(\sqrt{n}\hat{\pi}_n).$$

Conclusions

- Triangular structural models facilitate causal analysis via recursive conditioning, directed acyclic graph representation.
- Recursive conditional quantile models yield interpretable heterogeneous structural effects.
- Control variate methods offer computationally and statistically efficient strategies for estimating heterogeneous structural effects.
- Weighted average derivative methods offer a less restrictive strategy for estimation that offers potential for model diagnostics and testing.

Appendix 4
Risk, Choquet Portfolios and Quantile Regression

Joint work with Gib Bassett (UIC) and Gregory Kordas (Athens)

Outline

- Is there a useful role for pessimism in decision theory?
- A pessimistic theory of risk
- How to be pessimistic?

St. Petersburg Paradox

What would you be willing to pay to play the game:

$$G = \{\text{pay: } p, \quad \text{win: } 2^n \text{ with probability } 2^{-n}, \quad n = 1, 2, ...\}$$

Daniel Bernoulli ($\sim$ 1728) observed that even though the expected payoff was infinite, the gambler who maximized logarithmic utility would pay only a finite value to play. For example, given initial wealth 100,000 Roubles, our gambler would be willing to pay only 17 Roubles and 55 kopecks. If initial wealth were only 1000 Roubles, then the value of the game is only about 11 Roubles.

Expected Utility

To decide between two real valued gambles

$$X \sim F \quad \text{and} \quad Y \sim G$$

we choose X over Y if

$$Eu(X) = \int u(x)dF(x) \geqslant \int u(y)dG(y) = Eu(Y)$$

On Axiomatics

Suppose we have acts $P, Q, R, ...$ in a space $\mathcal{P}$, which admits enough convex structure to allow us to consider mixtures,

$$\alpha P + (1 - \alpha)Q \in \mathcal{P} \quad \alpha \in (0, 1)$$

Think of P, Q, R as probability measures on some underlying outcome/event space, $\mathcal{X}$.
Or better, view P, Q, R as acts mapping a space $\mathcal{S}$ of soon-to-be-revealed "states of nature" to the space of probability measures on the outcome space, $\mathcal{X}$.

The Expected Utility Theorem

Theorem(von-Neumann-Morgenstern) Suppose we have a preference relation $\{\succeq, \succ, \sim\}$ on $\mathcal{P}$ satisfying the axioms:

(A.1) (weak order) For all $P, Q, R \in \mathcal{P}$, $P \succeq Q$ or $Q \succeq P$, and $P \succeq Q$ and $Q \succeq R \Rightarrow P \succeq R$,

(A.2) (independence) For all $P, Q, R \in \mathcal{P}$ and $\alpha \in (0,1)$, then $P \succ Q \Rightarrow \alpha P + (1-\alpha)R \succ \alpha Q + (1-\alpha)R$,

(A.3) (continuity) For all $P, Q, R \in \mathcal{P}$, if $P \succ Q$ and $Q \succ R$, then there exist α and $\beta \in (0,1)$, such that, $\alpha P + (1-\alpha)R \succ \beta Q(1-\beta)R$.

Then there exists a linear function u on $\mathcal{P}$ such that for all $P, Q \in \mathcal{P}$, $P \succ Q$ if and only if $u(P) > u(Q)$.

Weakening the Independence Axiom

The independence axiom seems quite innocuous, but it is extremely powerful. We will consider a weaker form of independence due to Schmeidler (1989).

(A.2') (comonotonic independence) For all pairwise comonotonic $P, Q, R \in \mathcal{P}$ and $\alpha \in (0,1)$ $P \succ Q \Rightarrow \alpha P + (1-\alpha)R \succ \alpha Q + (1-\alpha)R$,

Definition Two acts P and Q in $\mathcal{P}$ are comonotonic, or similarly ordered, if for no s and t in $\mathcal{S}$,

$$P(\{t\}) \succ P(\{s\}) \quad \text{and} \quad Q(\{s\}) \succ Q(\{t\}).$$

"If P is better in state t than state s, then Q is also better in t than s."

On Comonotonicity

Definition The two functions $X, Y : \Omega \to \mathbb{R}$ are comonotonic if there exists a third function $Z : \Omega \to \Re$ and increasing functions f and g such that $X = f(Z)$ and $Y = g(Z)$.

From our point of view the crucial property of comonotonic random variables is the behavior of quantile functions of their sums. For comonotonic random variables X, Y, we have

$$F_{X+Y}^{-1}(u) = F_X^{-1}(u) + F_Y^{-1}(u)$$

By comonotonicity we have a $U \sim U[0,1]$ such that $Z = g(U) = F_X^{-1}(U) + F_Y^{-1}(U)$ where g is left continuous and increasing, so by monotone invariance, $F_{g(U)}^{-1} = g \circ F_U^{-1} = F_X^{-1} + F_Y^{-1}$. Comonotonic random variables are maximally dependent *a la* Fréchet

Choquet Expected Utility

Among the many proposals offered to extend expected utility theory the most attractive (to us) replaces

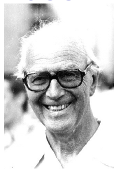
$$E_F u(X) = \int_0^1 u(F^{-1}(t))dt \geqslant \int_0^1 u(G^{-1}(t))dt = E_G u(Y)$$

with

$$E_{\nu,F} u(X) = \int_0^1 u(F^{-1}(t))d\nu(t) \geqslant \int_0^1 u(G^{-1}(t))d\nu(t) = E_{\nu,G} u(Y)$$

The measure ν permits distortion of the probability assessments after ordering the outcomes. This rank dependent form of expected utility has been pioneered by Quiggin (1981), Yaari (1987), Schmeidler (1989), Wakker (1989) and Dennenberg (1990).

Pessimism



By relaxing the independence axiom we obtain a larger class of preferences representable as Choquet capacities and introducing pessimism. The simplest form of Choquet expected utility is based on the “distortion”

$$\nu_\alpha(t) = \min\{t/\alpha, 1\}$$

so

$$E_{\nu_\alpha,F} u(X) = \alpha^{-1} \int_0^\alpha u(F^{-1}(t))dt$$

This exaggerates the probability of the proportion α of least favorable events, and totally discounts the probability of the $1 - \alpha$ most favorable events.

Expect the worst – and you won’t be disappointed.

A Smoother example

A simple, yet intriguing, one-parameter family of pessimistic Choquet distortions is the measure:

$$\nu_\theta(t) = 1 - (1-t)^\theta \qquad \theta \geqslant 1$$

Note that, changing variables, $t \to F_X(u)$, we have,

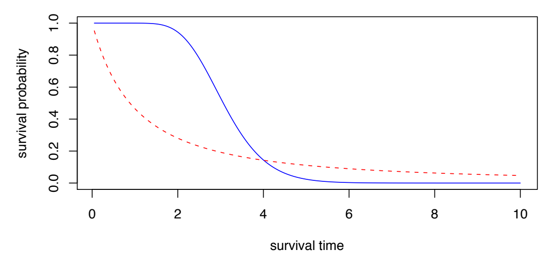
$$E_{\nu_\theta} X = \int_0^1 F_X^{-1}(t)d\nu(t) = \int_{-\infty}^{\infty} ud(1 - (1 - F_X(u))^\theta)$$

The pessimist imagines that he gets not a single draw from X but θ draws, and from these he always gets the worst. The parameter θ is a natural “measure of pessimism,” and need not be an integer.

Savage on Pessimism

I have, at least once heard it objected against the personalistic view of probability that, according to that view, two people might be of different opinions, according as one is pessimistic and the other optimistic. I am not sure what position I would take in abstract discussion of whether that alleged property of personalistic views would be objectionable, but I think it is clear from the formal definition of qualitative probability that the particular personalistic view sponsored here does not leave room for optimism and pessimism, however these traits may be interpreted, to play any role in the person’s judgement of probabilities. (Savage(1954), p. 68)

Pessimistic Medical Decision Making?



Survival Functions for a hypothetical medical treatment: The Lehmann quantile treatment effect (QTE) is the horizontal distance between the survival curves. In this example consideration of the mean treatment effect would slightly favor the (dotted) treatment curve, but the pessimistic patient might favor the (solid) placebo curve. Only the luckiest 15% actually do better under the treatment.

Risk as Pessimism?

In expected utility theory risk is entirely an attribute of the utility function:

Risk Neutrality	$\Rightarrow$	$u(x) \sim$ affine
Risk Aversion	$\Rightarrow$	$u(x) \sim$ concave
Risk Attraction	$\Rightarrow$	$u(x) \sim$ convex

Locally, the risk premium, i.e. the amount one is willing to pay to accept a zero mean risk, X, is

$$\pi(w, X) = \tfrac{1}{2}A(w)V(X)$$

where $A(w) = -u''(w)/u'(w)$ is the Arrow-Pratt coefficient of absolute risk aversion and $V(X)$ is the variance of X. Why is variance a reasonable measure of risk?

A Little Risk Aversion is a Dangerous Thing

Would you accept the gamble:

$$G_1 \qquad 50-50 \left\langle \begin{array}{l} \text{win } \$110 \\ \text{lose } \$100 \end{array} \right.$$

Suppose you say "no", then what about the gamble:

$$G_2 \qquad 50-50 \left\langle \begin{array}{l} \text{win } \$700,000 \\ \text{lose } \$1,000 \end{array} \right.$$

If you say "no" to G_1 for any initial wealth up to \$300,000, then you must also say "no" to G_2.
Moral: A little local risk aversion over small gambles implies implausibly large risk aversion over large gambles. Reference: Rabin (2000)

Are Swiss Bicycle Messengers Risk Averse?

When Veloblitz and Flash bicycle messengers from Zurich were confronted with the bet:

$$50-50 \left\langle \begin{array}{l} \text{win } 8 \ CHF \\ \text{lose } 5 \ CHF \end{array} \right.$$

More than half (54%) rejected the bet.
Reference: Fehr and Götte (2002)

Coherent Risk

Definition (Artzner, Delbaen, Eber and Heath (1999)) For real valued random variables $X \in \mathcal{X}$ on $(\Omega, \mathcal{A})$ a mapping $\rho : \mathcal{X} \to \mathcal{R}$ is called a coherent risk measure if,

1. Monotone: $X, Y \in \mathcal{X}$, with $X \leqslant Y \Rightarrow \rho(X) \geqslant \rho(Y)$.
2. Subadditive: $X, Y, X+Y \in \mathcal{X}, \Rightarrow \rho(X+Y) \leqslant \rho(X) + \rho(Y)$.
3. Linearly Homogeneous: For all $\lambda \geqslant 0$ and $X \in \mathcal{X}$, $\rho(\lambda X) = \lambda \rho(X)$.
4. Translation Invariant: For all $\lambda \in \mathcal{R}$ and $X \in \mathcal{X}$, $\rho(\lambda + X) = \rho(X) - \lambda$.

Many conventional measures of risks including those based on standard deviation are ruled out by these requirements. So are quantile based measures like "value at risk."

Choquet α-Risk

The leading example of a coherent risk measure is

$$\rho_{\nu_\alpha}(X) = -\alpha^{-1} \int_0^\alpha F^{-1}(t)dt$$

Variants of this risk measure have been introduced under several names

- Expected shortfall (Acerbi and Tasche (2002))
- Conditional *VaR* (Rockafellar and Uryasev (2000))
- Tail conditional expectation (Artzner, et al (1999)).

Note that $\rho_{\nu_\alpha}(X) = -E_{\nu_\alpha,F}(X)$, so Choquet α-risk is just negative Choquet expected utility with the distortion function ν_α.

Pessimistic Risk Measures

Definition A risk measure ρ will be called pessimistic if, for some probability measure φ on [0, 1]

$$\rho(X) = \int_0^1 \rho_{\nu_\alpha}(X)d\varphi(\alpha)$$

By Fubini

$$\begin{aligned}\rho(X) &= -\int_0^1 \alpha^{-1}\int_0^\alpha F^{-1}(t)dtd\varphi(\alpha)\\ &= -\int_0^1 F^{-1}(t)\int_t^1 \alpha^{-1}d\varphi(\alpha)dt\\ &\equiv -\int_0^1 F^{-1}(t)d\nu(t)\end{aligned}$$

Approximating General Pessimistic Risk Measures

We can approximate any pessimistic risk measure by taking

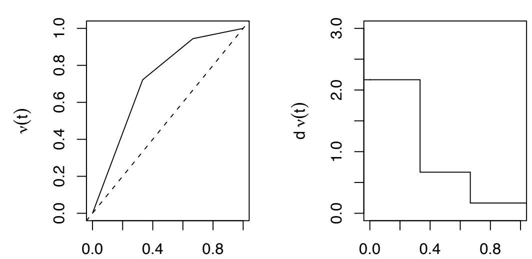
$$d\varphi(t) = \sum \varphi_i \delta_{\tau_i}(t)$$

where δ_τ denotes (Dirac) point mass 1 at τ. Then

$$\rho(X) = -\varphi_0 F^{-1}(0) - \int_0^1 F^{-1}(t)\gamma(t)dt$$

where $\gamma(t) = \sum \varphi_i \tau_i^{-1} I(t < \tau_i)$ and $\varphi_i > 0$, with $\sum \varphi_i = 1$.

An Example



$$d\varphi(t) = \tfrac{1}{2}\delta_{1/3}(t) + \tfrac{1}{3}\delta_{2/3}(t) + \tfrac{1}{6}\delta_1(t)$$

A Theorem

Theorem (Kusuoka (2001)) A regular risk measure is *coherent* in the sense of Artzner *et. al.* if and only if it is *pessimistic*.

- Pessimistic Choquet risk measures correspond to *concave* ν, i.e., *monotone decreasing* $d\nu$.
- Probability assessments are distorted to accentuate the probability of the least favorable events.
- The crucial coherence requirement is subadditivity, or submodularity, or 2-alternatingness in the terminology of Choquet capacities.

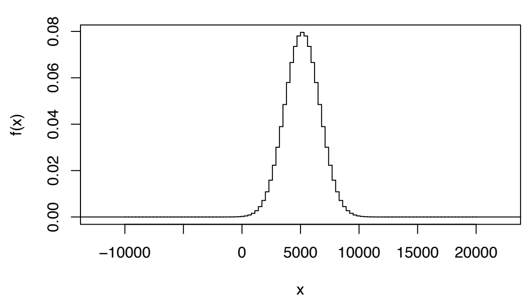
An Example

Samuelson (1963) describes asking a colleague at lunch whether he would be willing to make a

$$50-50 \text{ bet} \left\langle \begin{array}{l} \text{win } 200 \\ \text{lose } 100 \end{array} \right.$$

The colleague (later revealed to be E. Cary Brown) responded
"no, but I *would* be willing to make 100 such bets."
This response has been interpreted not only as reflecting a basic confusion about how to maximize expected utility but also as a fundamental misunderstanding of the law of large numbers.

Payoff Density of 100 Samuelson trials



Odds of losing money on the 100 trial bet is 1 chance in 2300.

Was Brown really irrational?

Suppose, for the sake of simplicity that

$$d\varphi(t) = \tfrac{1}{2}\delta_{1/2}(t) + \tfrac{1}{2}\delta_1(t)$$

so for one Samuelson coin flip we have the unfavorable evaluation,

$$E_{\nu,F}(X) = \tfrac{1}{2}(-100) + \tfrac{1}{2}(50) = -25$$

but for $S = \sum_{i=1}^{100} X_i \sim \mathcal{B}in(.5, 100)$ we have the favorable evaluation,

$$\begin{aligned} E_{\nu,F}(S) &= \tfrac{1}{2}2\int_0^{1/2} F_S^{-1}(t)dt + \tfrac{1}{2}(5000) \\ &= 1704.11 + 2500 \\ &= 4204.11 \end{aligned}$$

How to be Pessimistic

Theorem Let X be a real-valued random variable with $EX = \mu < \infty$, and $\rho_\alpha(u) = u(\alpha - I(u < 0))$. Then

$$\min_{\xi \in \mathcal{R}} E\rho_\alpha(X - \xi) = \alpha\mu + \rho_{\nu_\alpha}(X)$$

So α risk can be estimated by the sample analogue

$$\hat{\rho}_{\nu_\alpha}(x) = (n\alpha)^{-1} \min_{\xi} \sum \rho_\alpha(x_i - \xi) - \hat{\mu}_n$$


I knew it! Eventually everything looks like quantile regression to this guy!

Pessimistic Portfolios

Now let $X = (X_1, \dots, X_p)$ denote a vector of potential portfolio asset returns and $Y = X^\top \pi$, the returns on the portfolio with weights π. Consider

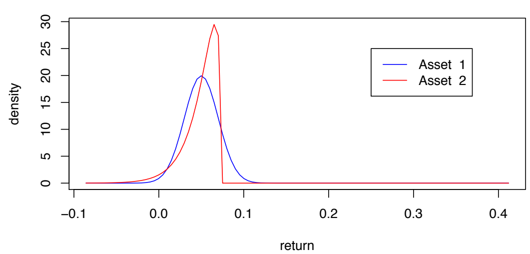
$$\min_{\pi} \rho_{\nu_\alpha}(Y) - \lambda\mu(Y)$$

Minimize α-risk subject to a constraint on mean return.
This problem can be formulated as a linear quantile regression problem

$$\min_{(\beta,\xi)\in\mathcal{R}^p} \sum_{i=1}^{n} \rho_\alpha(x_{i1} - \sum_{j=2}^{p}(x_{i1} - x_{ij})\beta_j - \xi) \quad s.t. \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

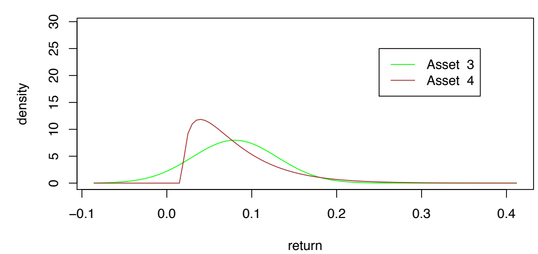
where $\pi(\beta) = (1 - \sum_{j=2}^{p} \beta_j, \beta^\top)^\top$.

An Example



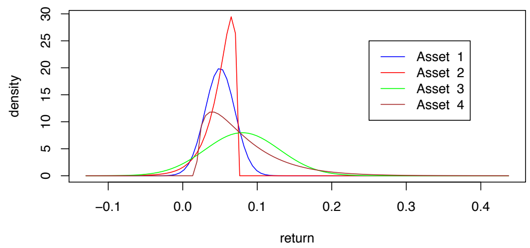
Two asset return densities with identical mean and variance.

An Example



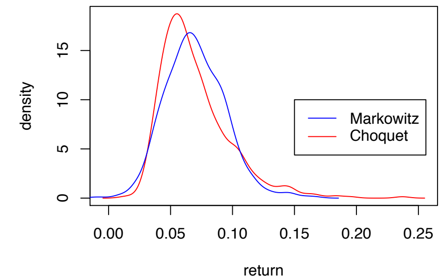
Two more asset return densities with identical mean and variance.

An Example



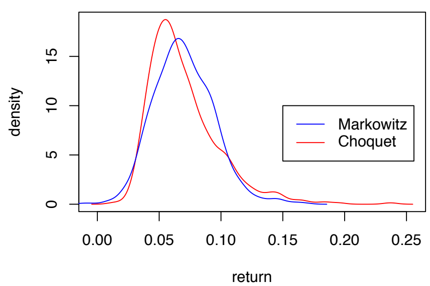
Two pairs of asset return densities with identical mean and variance.

Optimal Choquet and Markowitz Portfolio Returns



Markowitz portfolio minimizes the standard deviation of returns subject to mean return $\mu = .07$. The Choquet portfolio minimizes Choquet risk (for $\alpha = .10$) subject to earning the same mean return. The Choquet portfolio has better performance in both tails than mean-variance Markowitz portfolio.

Optimal Choquet and Markowitz Portfolio Returns



Now, the Markowitz portfolio minimizes the standard deviation of returns subject to mean return $\mu = .07$. The Choquet portfolio maximizes expected return subject to achieving the same Choquet risk (for $\alpha = .10$) as the Markowitz portfolio. Choquet portfolio has expected return $\mu = .08$ a full percentage point higher than the Markowitz portfolio.

A Unified Theory of Pessimism

Any pessimistic risk measure may be approximated by

$$\rho_\nu(X) = \sum_{k=1}^{m} \varphi_k \rho_{\nu_{\alpha_k}}(X)$$

where $\varphi_k > 0$ for $k = 1, 2, ..., m$ and $\sum \varphi_k = 1$.
Portfolio weights can be estimated for these risk measures by solving linear programs that are weighted sums of quantile regression problems:

$$\min_{(\beta,\xi)\in\Re^p} \sum_{k=1}^{m} \sum_{i=1}^{n} \nu_k \rho_{\alpha_k}(x_{i1} - \sum_{j=2}^{p}(x_{i1} - x_{ij})\beta_j - \xi_k) \quad s.t. \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

Software in R is available on from my web pages.

Conclusions

- Expected Utility is unsatisfactory both as a positive, i.e., descriptive, theory of behavior and as a normative guide to behavior.
- Choquet (non-additive, rank dependent) expected utility provides a simple, tractable alternative.
- Mean-variance Portfolio allocation is also unsatisfactory since it relies on unpalatable assumptions of Gaussian returns, or quadratic utility.
- Choquet portfolio optimization can be formulated as a quantile regression problem thus providing an attractive practical alternative to the dominant mean-variance approach of Markowitz (1952).

Appendix 5: Yet Another R FAQ, or How I Learned to Stop Worrying and Love Computing

More official R FAQs are available from the CRAN website. A FAQ for the quantile regression package **quantreg** can be found by the invoking the command `FAQ()` from within R after loading the package.

"It was a splendid mind. For if thought is like the keyboard of a piano, divided into so many notes, or like the alphabet is ranged in twenty-six letters all in order, then his splendid mind had no sort of difficulty in running over those letters one by one, firmly and accurately, until it had reached the letter Q. He reached Q. Very few people in the whole of England reach the letter Q.... But after Q? What comes next?... Still, if he could reach R it would be something. Here at least was Q. He dug his heels in at Q. Q he was sure of. Q he could demonstrate. If Q then is Q–R–.... Then R... He braced himself. He clenched himself.... In that flash of darkness he heard people saying–he was a failure–that R was beyond him. He would never reach R. On to R, once more. R—.... ...He had not genius; he had no claim to that: but he had, or he might have had, the power to repeat every letter of the alphabet from A to Z accurately in order. Meanwhile, he stuck at Q. On then, on to R."

Virginia Woolf (To the Lighthouse)

R FAQ

1. How to get it? Google CRAN, click on your OS, and download. Buy a case of wine with what you've saved.
2. How to start? Click on the R icon if you are mousey, type R in a terminal window if you are penguinesque.
3. What next? At the prompt, $>$ type `2 + 2`
4. What next? At the prompt, $>$ type `1:9/10`
5. What next? At the prompt, $>$ type `x <- 1:99/100`
6. What next? At the prompt, $>$ type `plot(x,sin(1/x))`
7. What next? At the prompt, $>$ type `lines(x,sin(1/x),col = "red")`
8. How to stop? Click on the Stop sign if you are mousey, type `q()` if you are penguinesque.

R FAQ

9. Isn't there more to R? Yes, try downloading some packages: using the menu in the GUI if you are mousey, or typing install.packages("pname") if you are penguinesque.
10. What's a package? A package is a collection of R software that augments in some way the basic functionality of R, that is it is a way of going "beyond R." For example, the **quantreg** package is a collection of functions to do quantile regression. There were 2992 packages on CRAN as of July 9, 2014.
11. How to use a package? Downloading and installing a package isn't enough, you need to tell R that you would like to use it, for this you can either type: `require(pname)` or `library(pname)`.

R FAQ

12. How to read data files? For simple files with values separated by white space you can use `read.table`, or `read.csv` for data separated by commas, or some other mark. For more exotic files, there is `scan`. And for data files from other statistical environments, there is the package **foreign** which facilitates the reading of Stata, SAS and other data. There are also very useful packages to read html and other files from the web, but this takes us beyond our introductory objective.
13. What is a `data.frame`? A `data.frame` is a collection of related variables; in the simplest case it is simply a data matrix with each row indexing an observation. However, unlike conventional matrices, the columns of a `data.frame` can be non-numeric, e.g. logical or character or in R parlance, "factors." In many R functions one can specify a `data = "dframe"` argument that specifies where to find the variables mentioned elsewhere in the call.

R FAQ

14. How to get help? If you know what command you want to use, but need further details about how to use it, you can get help by typing `?fname`, if you don't know the function name, then you might try `apropos("concept")`. If this fails then a good strategy is to search `http://finzi.psych.upenn.edu/search.html` with some relevant keywords; here you can specify that you would like to search through the R-help newsgroup, which is a rich source of advice about all things R.
15. Are there manuals? Yes, of course there are manuals, but only to be read as a last resort, but when things get desparate you can always RTFM. The left side of the CRAN website has links to manuals , FAQs and contributed documentation. Some of the latter category is quite good, and is also available in a variety of natural languages. There is also an extensive shelf of published material about R, but indulging in this tends to put a crimp in one's wine budget.

16. What about illustrative examples? A strength of R is the fact that most of the documentation files for R functions have example code that can be easily executed. Thus, for example if you would like to see an example of how to use the command `rq` in the **quantreg** package, you can type `example(rq)` and you will see some examples of its use. Alternatively, you can cut and paste bits of the documentation into the R window; in the OSX GUI you can simply highlight code in a help document, or other window and then press Command-Enter to execute. Similarly, many packages have demo files that act as auxiliary documentation. To see what demos are available for currently loaded packages, just try `demos()`. Finally, many packages have vignettes, short overviews of various aspects of the functionality of the package usually with explicit examples of how to do things. For example, the **quantreg** package has three vignettes: one basic, one about survival modeling, and one about additive nonparametric models. Vignettes can be accessed from R by simply typing `vignette("vname")`. The names of the various package vignettes can be found by typing `vignette()`.

R FAQ

17. What's in a name? Objects in R can be of various types called classes. You can create objects by assignment, typically as above with a command like `f <- function(x,y,z)`. A list of the objects currently in your private environment can be viewed with `ls()`, objects in lower level environments like those of the packages that you have loaded can be viewed with `ls(k)` where `k` designates the number of the environment. A list of these environments can be seen with `search()`. Objects can be viewed by simply typing their name, but sometimes objects can be very complicated so a useful abbreviated summary can be obtained with `str(object)`.

R FAQ

18. What about my beloved least squares? Fitting linear models in R is like taking a breath of fresh air after inhaling the smog of other industrial environments. To do so, you specify a model `formula` like this: `lm(y ~ x1 + x2 + x3, data = "dframe")`, if one or more of the x's are factor variables, that is take discrete, qualitative values, then they are automatically exanded into several indicator variables. Interactions plus main effects can be specified by replacing the "+" in the formula by "*". Generalized linear models can be specified in much the same way, as can quantile regression models using the **quantreg** package.

R FAQ

19. What about class conflict? Class analysis can get complicated, but you can generally expect that classes behave themselves in accordance with their material conditions. Thus, for example, suppose you have fitted a linear regression model by least squares using the command `f <- lm(y ~ x1 + x2 + x3)`, thereby assigning the fitted object to the symbol `f`. The object `f` will have class `lm`, and when you invoke the command `summary(f)`, R will try to find a summary method appropriate to objects of class `lm`. In the simplest case this will entail finding the command `summary.lm` which will produce a conventional table of coefficients, standard errors, t-statistics, p-values and other descriptive statistics. Invoking `summary` on a different type of object, say a data.frame, will produce a different type of summary object. Methods for prediction, testing, plotting and other functionalities are also provided on a class specific basis.

20. What about graphics? R has a very extensive graphics capability. Interactive graphics of the type illustrated already above is quite simple and easy to use. For publication quality graphics, there are device drivers for various graphical formats, generally I find that `pdf` is satisfactory. Dynamic and 3D graphics can be accessed from the package **rgl**.
21. Latex tables? The packages **Hmisc** and **xtable** have very convenient functions to convert R matrices into latex tables.
22. Random numbers? There is an extensive capability for generating pseudo random numbers from R. Reproducibility of random sequences is ensured by using the `set.seed` command. Various distributions are accessible with families of functions using the prefixes pdqr, thus for example `pnorm`, `dnorm`, `qnorm` and `rnorm` can be used to evaluate the distribution function, density function, quantile function, or to generate random normals, respectively. See `?Distributions` for a complete list of standard distributions available in base R in this form. Special packages provide additional scope, although it is sometimes tricky to find them.

R FAQ

23. Programming and simulation? The usual language constructs for looping, switching and data management are available, as are recent developments for exploiting multicore parallel processing. Particularly convenient are the family of `apply` functions that facilitate summarizing matrix and list objects. A good way to learn the R language is to look at the code for existing functions. Most of this code is easily accessible from the R command line. If you simply type the name of an R function, you will usually be able to see its code on the screen. Sometimes of course, this code will involve calls to lower level languages, and this code would have to be examined in the source files of the system. But everything is eventually accessible. If you don't like the way a function works you can define a modified version of it for your private use. If you are inspired to write lower level code this is also easily incorporated into the language as explained in the manual called "Writing R Extensions."

References

Abrevaya, J. (2001). The effect of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26:247–259.

Abrevaya, J. and Dahl, C. M. (2008). The effects of birth inputs on birthweight. *Journal of Business and Economic Statistics*, 26:379–397.

Bang, H. and Tsiatis, A. a. (2002). Median regression with censored cost data. *Biometrics*, 58(3):643–9.

Belloni, A. and Chernozhukov, V. (2009). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39:82–130.

Bose, A. and S., C. (2003). Generalized bootstrap for estimators of minimizers of convex functions. *Journal of Statistical Planning and Inference*, 117:225–239.

Brown, B. M. and Wang, Y. G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 92:149–158.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18:5–46.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81:2205–2268.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 114:293–346.

Choi, H. M. and Hobert, J. P. (2013). Analysis of MCMC algorithms for bayesian linear regression with Laplace errors. *Journal of Multivariate Analysis*, 117:32–40.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2:267–277.

Doksum, K. and Gasko, M. (1990). On a correspondence between models in binary regression and survival analysis. *Intl. Stat. Rev.*, 58:243–252.

Dunson, T. and Taylor, J. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, 17(3):385–400.

Feng, X., He, X., and Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika*, 98:995–999.

Fenske, N., Kneib, T., and Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, 106:494–510.

Fitzenberger, B. (1997). Computational aspects of censored quantile regression. In Dodge, Y., editor, *L1-Statistical Procedures and Related Topics*. IMS Lecture Notes- Monograph Series.

Gelfand, A. E. and Kottas, A. (2002). A computational approach for full nonparametric bayesian inference under dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11(2):289–305.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8:140–154.

Geraci, M. and Bottai, M. (2013). Linear quantile mixed models. *Stat Comput*, DOI 10.1007/s11222-013-9381-9.

Goh, S. C. and Knight, K. (2009). Nonstandard quantile-regression inference. *Econometric Theory*, 25:1415–1432.

Gutenbrunner, C. and Jurečková, J. (1992). Regression quantile and regression rank score process in the linear model and derived statistics. *Ann. Statist.*, 20:305–330.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, 2:307–333.

He, X. and Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, 97:783–795.

Jung, S. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91(433):251–257.

Kim, M.-O. and Yang, Y. (2011). Semiparametric approach to a random effects quantile regression model. *Journal of the American Statistical Association*, 106(496):1405–1417.

Kocherginsky, M., He, X., and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14:41–55.

Koenker, R. (1994). Confidence intervals for regression quantiles, in P. Mandl and M. Huskova (eds.). *Asymptotic Statistics*, pages 349–359.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

Koenker, R. (2008). Censored quantile regression redux. *Journal of Statistical Software*, 27.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and Hallock, K. (2000). Quantile regression: An introduction. `www.econ.uiuc.edu/~roger/research/intro/intro.html`.

Koenker, R. and Park, B. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71:265–283.

Kottas, A. and Krnjajić, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81:1565–1578.

Lancaster, T. and Jun, S. J. (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25(2):287–307.

Leng, C. and Zhang, W. (2014). Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing*, 24:123–136.

Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556.

Lum, K. and Gelfand, A. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, 7:235–258.

Machado, J. and Mata, J. (2001). Counterfactual decomposition of changes in wage distributions using quantile regression. *EE*, 26:115–134.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *J. of Eonometrics*, 3:205–228.

Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *The Annals of Statistics*, 28:1083–1104.

Mosteller, F. and Tukey, J. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading Mass.

Otsu, T. (2008). Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics*, 142(1):508–538.

Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC Press.

Parzen, M. I., Wei, L., and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, 81:341–350.

Peng, L. (2012). Self-consistent estimation of censored quantile regression. *Journal of Multivariate Analysis*, 105:368–379.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, 32:143–155.

Reich, B. J., Bondell, H. D., and Wang, H. J. (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, 11(2):337–352.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association*, 106:6–20.

Rosenblatt, F. (1958). The perceptron: a probablistic model for information storage and retrieval in the brain. *Psychological Review*, 65:386–408.

Stengos, T. and Wang, D. (2007). An algorithm for censored quantile regressions. *Economics Bulletin*, 3:1–9.

Tang, Y., Wang, H., He, X., and Zhu, Z. (2012). Informative subset estimation for censored quantile regression. *TEST*, 21:635–655.

Tokdar, S. T. and Kadane, J. B. (2011). Simulatenous linear quantile regression: A semiparametric bayesian approach. *Bayesian Analysis*, 6:1–22.

Tsionas, E. G. (2003). Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, 73:659–674.

Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of california boys and girls from birth to age 18. *California Public Child Development*, 1:183–364.

Wang, H. and He, X. (2007). Detecting differential expressions in genechip microarray studies. *Journal of the American Statistical Association*, 102(477):104–112.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 104(487):1117–1128.

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25:1369–1382.

Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, 40:1102–1131.

Yang, Y., Wang, H., and He, X. (2014). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *Technical report*.

Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association*, 90:178–184.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics and Probability Letters*, 54:437–447.

Zhou, L. (2006). A simple censored median regression estimator. *Statistica Sinica*, 16:1043–1058.