A Short Course on Quantile Regression

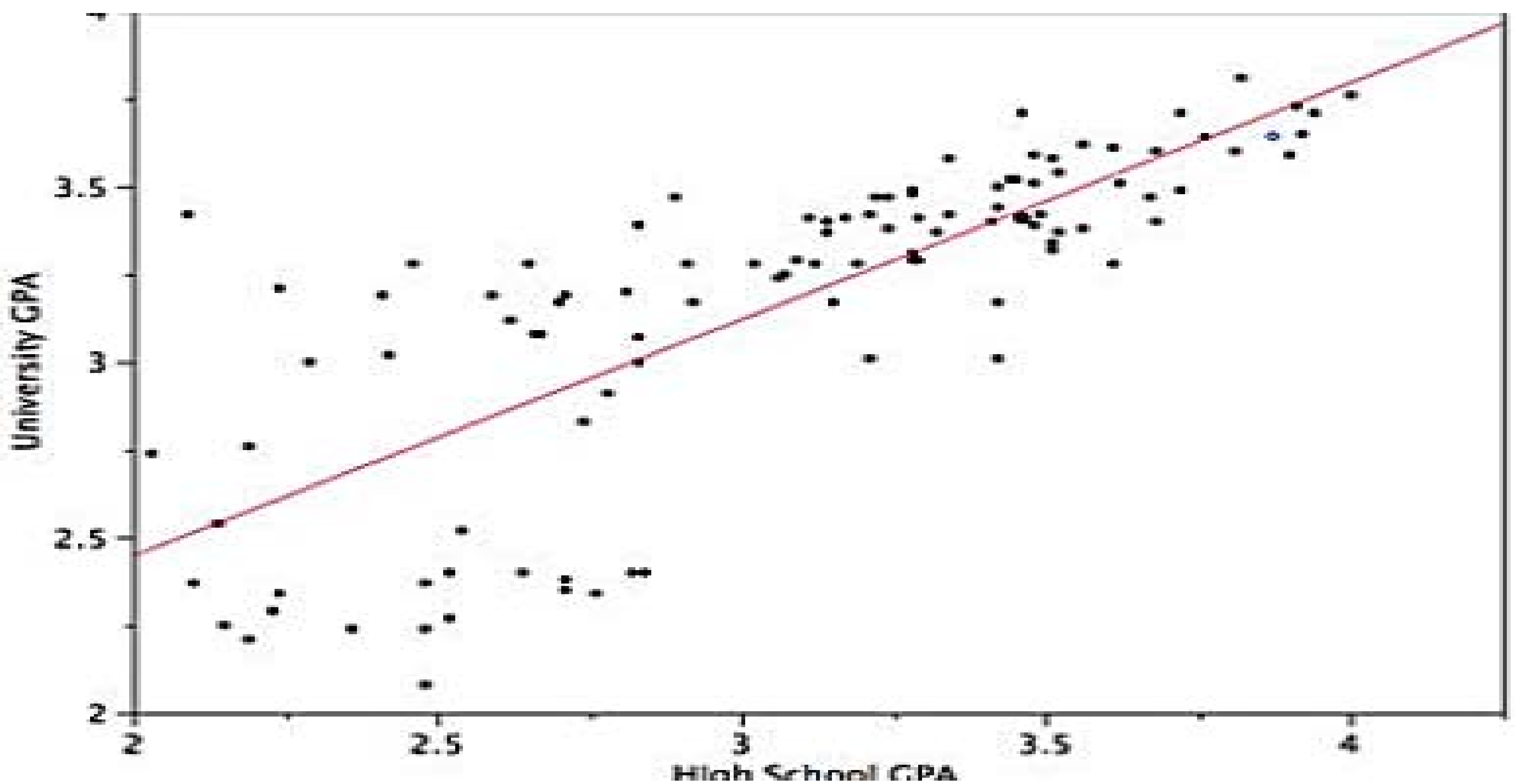
Xuming He (University of Michigan)

Course Outline:

1. Introduction to quantile regression
2. Basic properties of quantile regression estimates
3. Inference on quantile regression
4. Algorithm by linear programming; computer code by R and SAS
5. Examples
6. Nonparametric quantile curves
7. Censored quantile regression
8. Applications

1 Introduction to Quantile Regression

1.1 What is regression? (College GPA versus High School GPA)



1.2 Quantile Regression versus Mean Regression

Quantile. Let Y be a random variable with cumulative distribution function CDF $F_Y(y) = P(Y \leq y)$. The **τth quantile** of Y is

$$Q_\tau(Y) = \inf\{y : F_Y(y) \geq \tau\},$$

where $0 < \tau < 1$ is the quantile level.

- $Q_{0.5}(Y)$: median, the second quartile
- $Q_{0.25}(Y)$: the first quartile, 25th percentile
- $Q_{0.75}(Y)$: the third quartile, 75th percentile

Note: $Q_\tau(Y)$ is a **nondecreasing function** of τ, i.e.
$Q_{\tau_1}(Y) \leq Q_{\tau_2}(Y)$ for $\tau_1 < \tau_2$.

Conditional quantile. Suppose Y is the response variable, and $\mathbf{X}$ is the p-dimensional predictor. Let $F_Y(y|\mathbf{X}=\mathbf{x}) = P(Y \le y|\mathbf{X}=\mathbf{x})$ denote the conditional CDF of Y given $\mathbf{X}=\mathbf{x}$. Then the **τth conditional quantile** of Y is defined as

$$Q_\tau(Y|\mathbf{X}=\mathbf{x}) = \inf\{y : F_Y(y|\mathbf{x}) \ge \tau\}.$$

Least squares linear (mean) regression model :

$$Y = \mathbf{X}^T\boldsymbol{\beta} + U, \quad E(U) = 0.$$

Thus

$$E(Y|\mathbf{X}=\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ measures the marginal change in the mean of Y due to a marginal change in $\mathbf{x}$.

Linear quantile regression model:

$$Q_\tau(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}(\tau), \quad 0 < \tau < 1,$$

where $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \cdots, \beta_p(\tau))^T$ is the quantile coefficient that may depend on τ;

- the first element of $\mathbf{x}$ is one corresponding to the intercept, i.e. $\mathbf{x} = (1, x_2, \cdots, x_p)^T$;
- so that $Q_\tau(Y|\mathbf{x}) = \beta_1(\tau) + x_2\beta_2(\tau) + \cdots + x_p\beta_p(\tau)$;
- $\boldsymbol{\beta}(\tau)$ is the marginal change in the τth quantile due to the marginal change in $\mathbf{x}$.

Note that $Q_\tau(Y|\mathbf{x})$ is a nondecreasing function of τ for any given $\mathbf{x}$.

Example: location-scale shift model

$$Y_i = \beta_1 + \beta_2 Z_i + (1 + \gamma Z_i)\epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} F(\cdot).$$

The conditional quantile function

$$Q_\tau(Y|\mathbf{X}_i) = \beta_1(\tau) + \beta_2(\tau) Z_i,$$

where

- $\mathbf{X}_i = (1, Z_i)^T$;
- $\beta_1(\tau) = \beta_1 + F^{-1}(\tau)$ is nondecreasing in τ;
- $\boldsymbol{\beta}_2(\tau) = \beta_2 + \gamma F^{-1}(\tau)$ may depend on τ. That is, the covariate is allowed to have a different impact on different quantiles of the Y distribution.

Location-shift model: $\boldsymbol{\gamma} = 0$, so that $\beta_2(\tau) = \beta_2$ is constant across quantile levels.

1.3 Quantile Treatment Effect

Quantile Treatment Effect

- $Z_i = 0$: control; $Z_i = 1$: treatment
- $Y_i|Z_i = 0 \sim F$ (control distribution) and $Y_i|Z_i = 1 \sim G$ (treatment distribution)
- Mean treatment effect:

$$\Delta = E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) = \int ydG(y) - \int ydF(y).$$

- Quantile treatment effect:

$$\delta(\tau) = Q_\tau(Y|Z_i = 1) - Q_\tau(Y|Z_i = 0) = G^{-1}(\tau) - F^{-1}(\tau).$$

- Thus

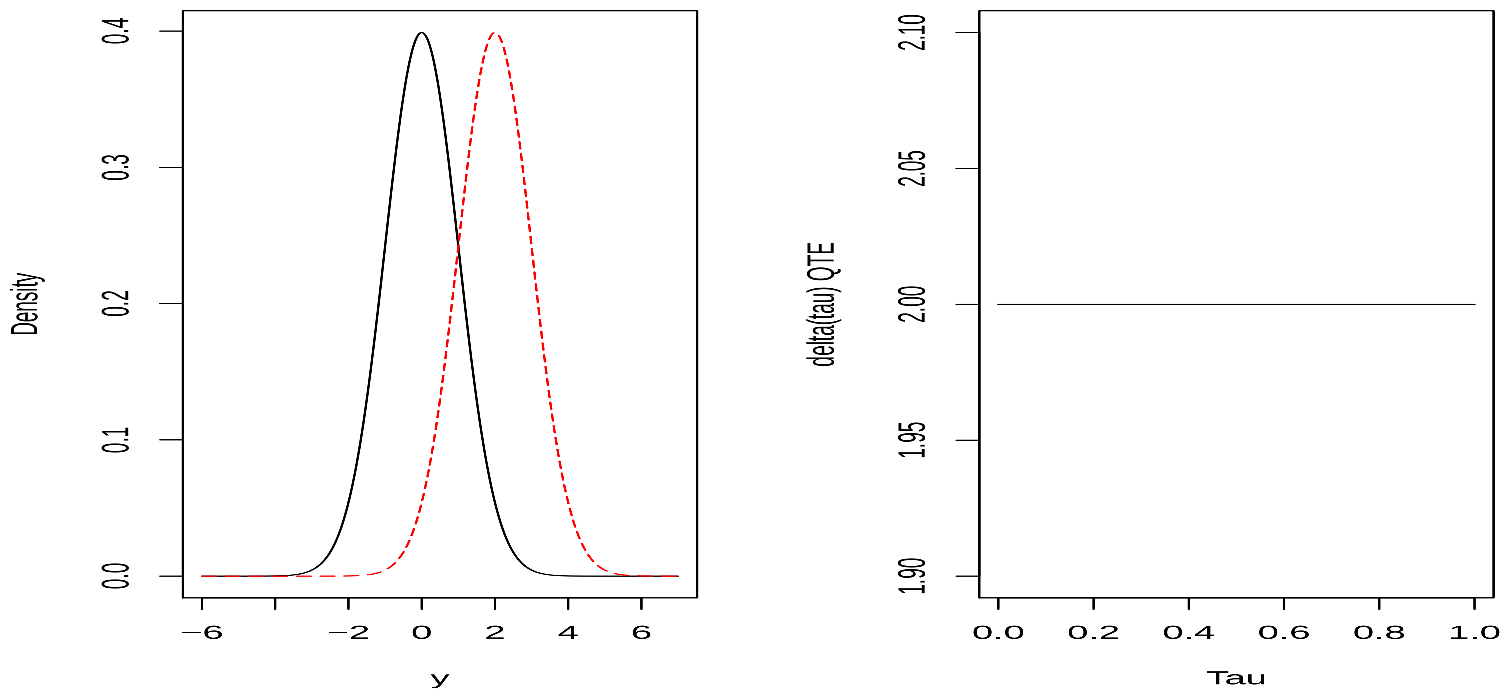
$$\Delta = \int_0^1 G^{-1}(u)du - \int_0^1 F^{-1}(u)du = \int_0^1 \delta(u)du.$$

- Equivalent quantile regression model (with binary covariate):

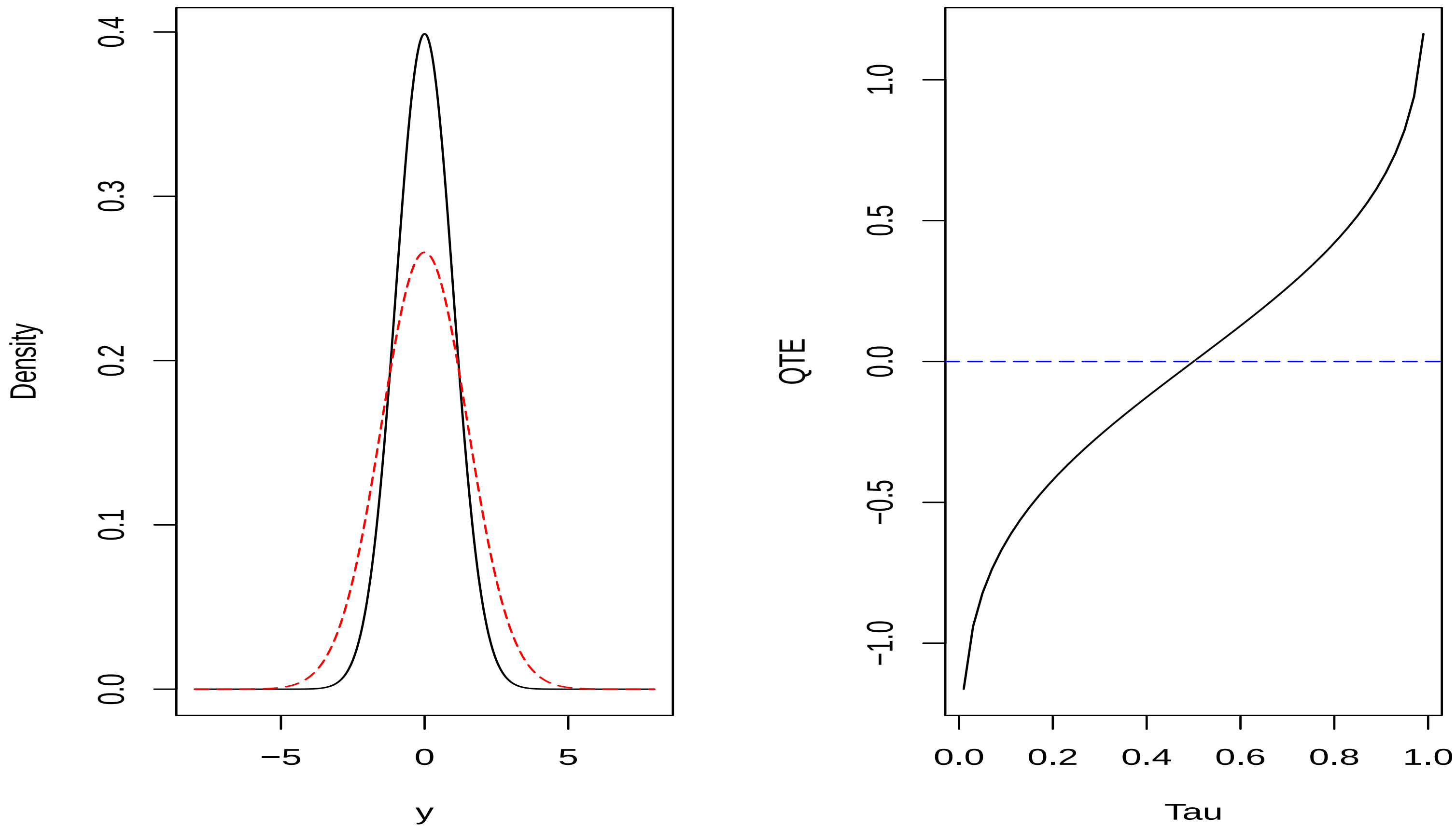
$$Q_\tau(Y|Z) = \alpha(\tau) + \delta(\tau)Z.$$

- **Location shift**:

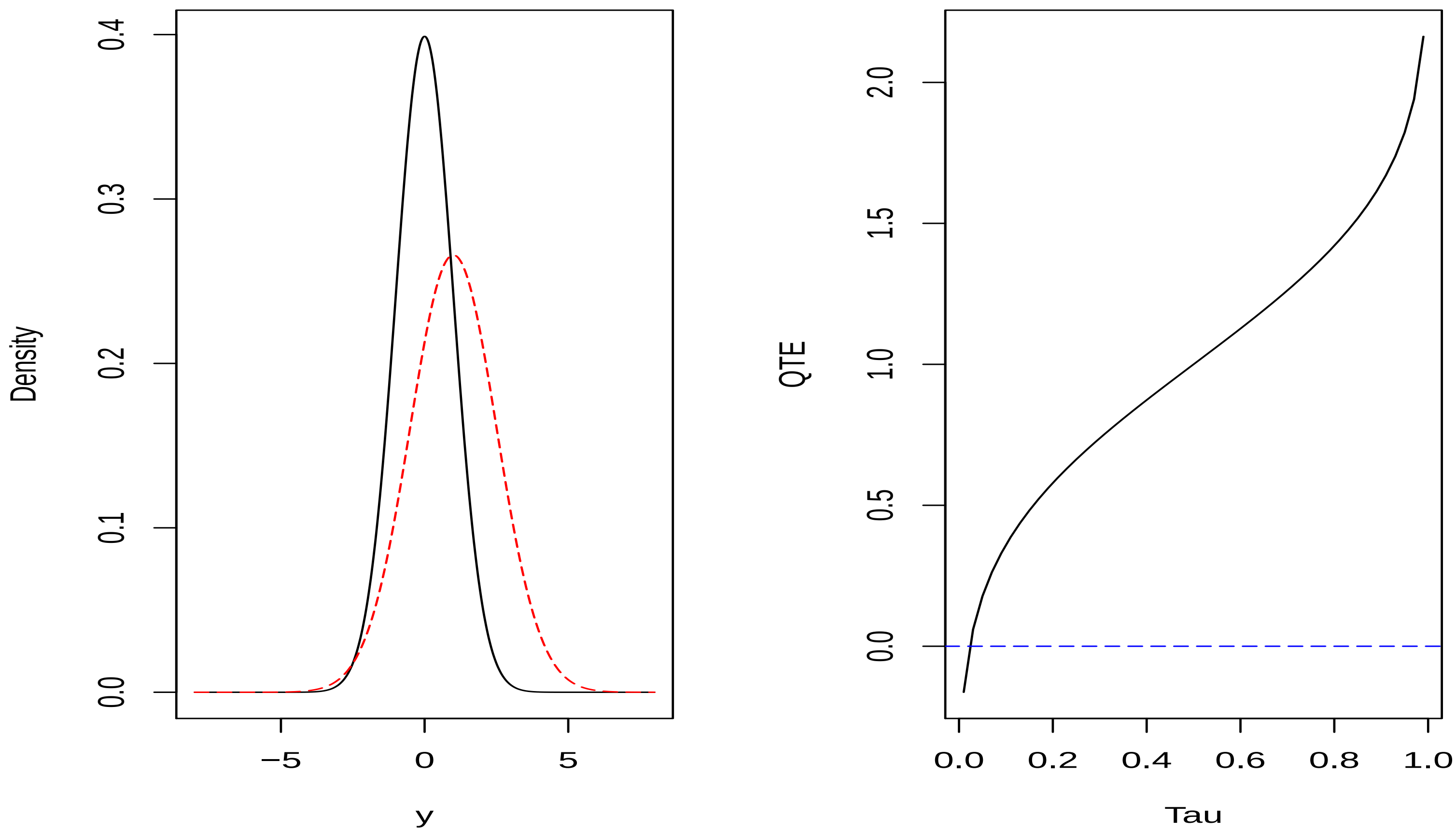
$$F(y) = G(y + \delta) \Rightarrow \delta(\tau) = \Delta = \delta.$$



- **Scale shift**: $\Delta = \delta(0.5) = 0$, but $\delta(\tau) \neq 0$ at other quantiles.



- **Location-scale shift**



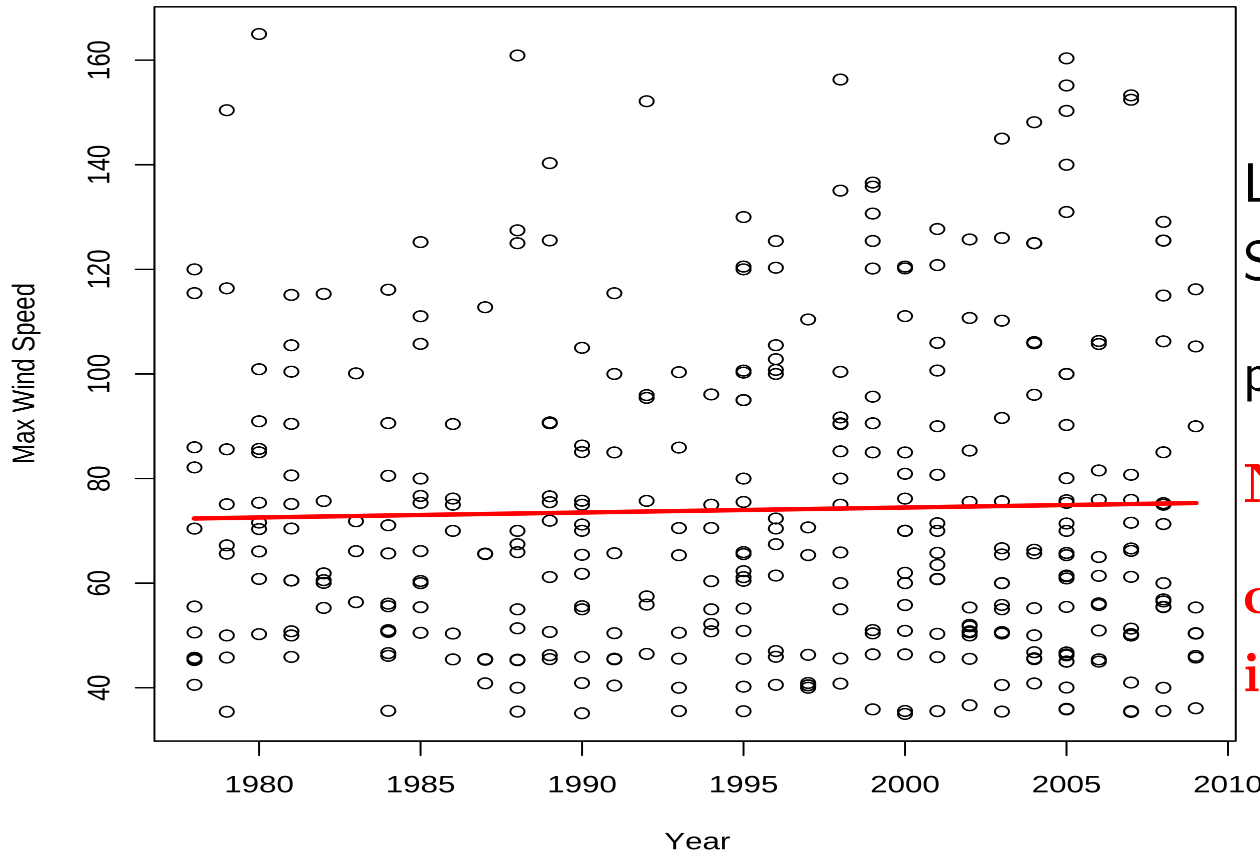
1.4 Advantages of Quantile Regression

Why Quantile Regression?

Case 1: Quantile regression allows us to study the impact of predictors on different quantiles of the response distribution, and thus provides a complete picture of the relationship between Y and $\mathbf{X}$.

Example: More Severe Tropical Cyclones?

- Y_i : max wind speeds of tropical cyclones in North Atlantic
- X_i: year 1978-2009



LS estimate

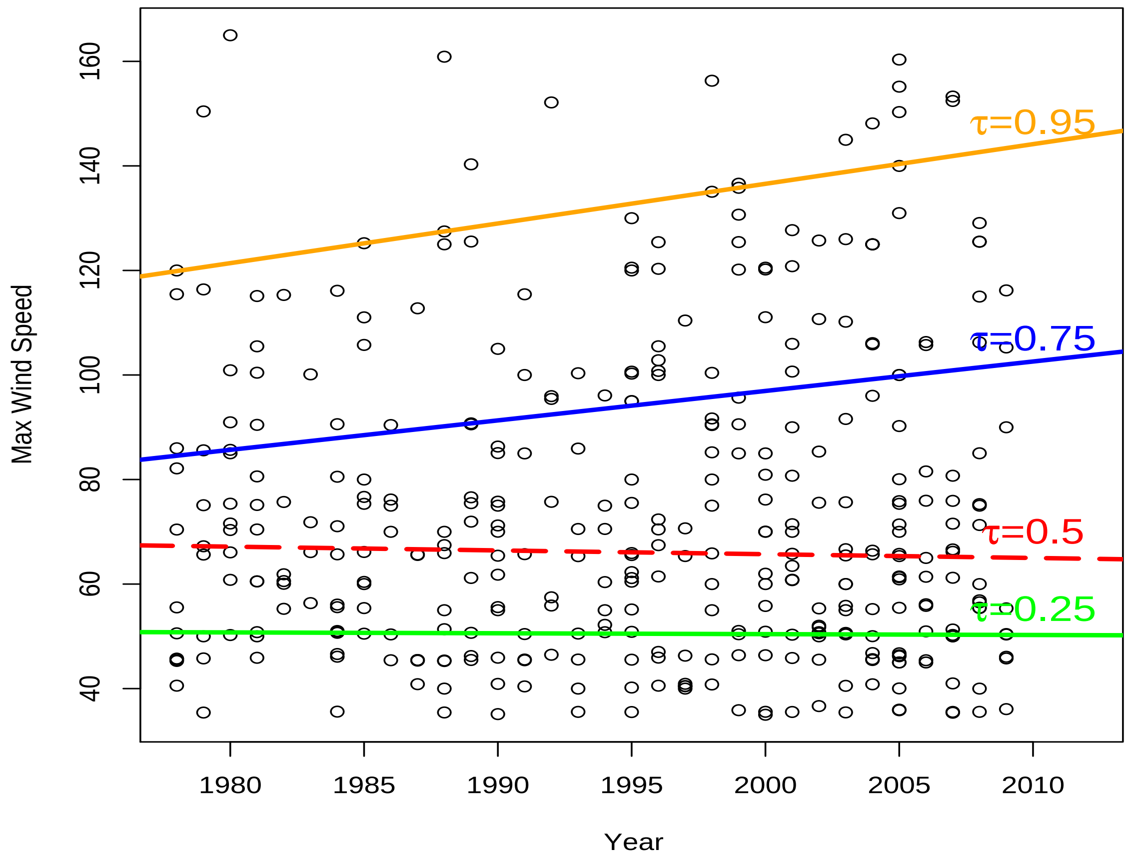
Slope: 0.095

p-val: 0.569

No significant trend in mean!

Q: Do the **quantiles** of max wind speed change over time?

τth quantile: $Q_\tau(Y) = \{y : P(Y < y) = \tau\}$.



p-value

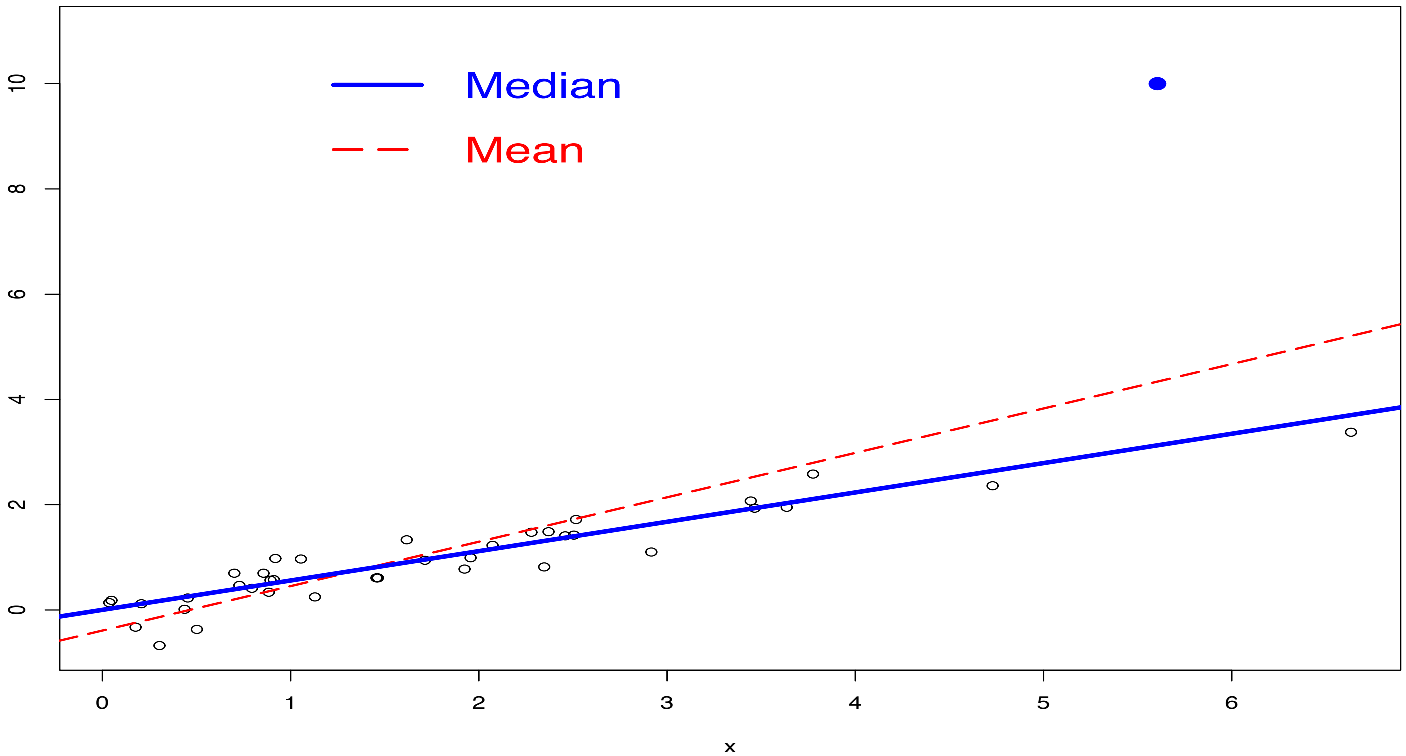
$\tau = 0.95$: 0.009

$\tau = 0.75$: 0.100

$\tau = 0.5$: 0.718

$\tau = 0.25$: 0.659

Case 2: robust to outliers in y observations.



Case 3: estimation and inference are distribution-free.

1.5 Estimation

Suppose we observe a random sample $\{y_i, \mathbf{x}_i, i = 1, \cdots, n\}$ of $(Y, \mathbf{X})$.

Mean and Least Squares Estimation (LSE)

- $E(Y) = \mu_Y = \arg\min_a E\{(Y - a)^2\}$.
- Sample mean solves $\min_a \sum_{i=1}^n (y_i - a)^2$.
- The least squares $\sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 = minimum$ is consistent for conditional mean $E(y|x) = \mathbf{x}^T \beta$.

Median and Least Absolute Deviation (LAD)

- Median $Q_{0.5}(Y) = \arg\min_a E|Y - a|$
- Sample median solves $\min_a \sum_{i=1}^n |y_i - a|$.
- Assume $med(y|x) = x^T\beta(0.5)$, then $\hat{\beta}(0.5)$ can be obtained by solving

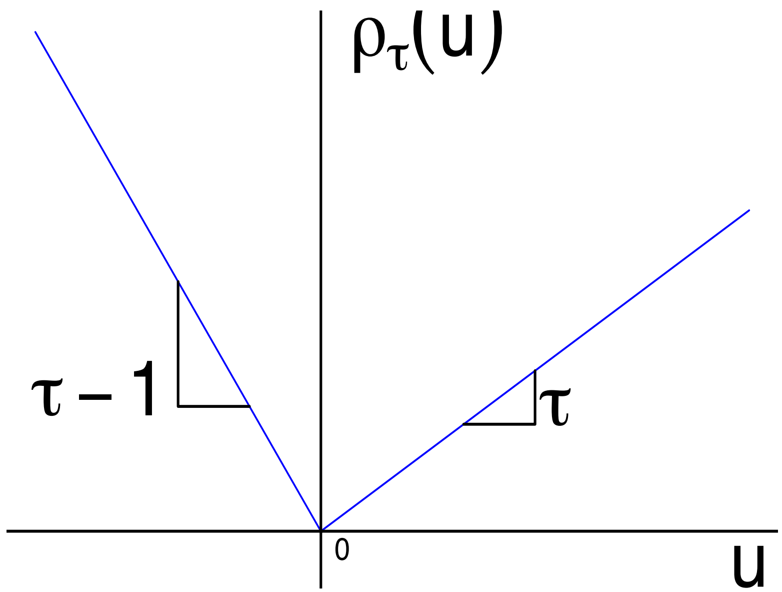
$$\min_\beta \sum_{i=1}^n |y_i - \mathbf{x}_i^T\beta|.$$

Quantile Regression at quantile level $0 < \tau < 1$

- τth quantile of Y:

$$Q_\tau(Y) = \arg\min_a E\{\rho_\tau(Y - a)\},$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function.



- τth sample quantile of Y solves

$$\min_a \sum_{i=1}^{n} \rho_\tau(y_i - a).$$

How to verify? Look at the gradient of the objective function as a function of a:

$$\tau \sum_i I(y_i - a > 0) = (1-\tau) \sum_i I(y_i - a < 0).$$

- Assume $Q_\tau(Y|\mathbf{X}) = \mathbf{X}^T \boldsymbol{\beta}(\tau)$, then

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

2 Basic Properties of Quantiles and Quantile Regression

2.1 Linear Programming (LP)

Linear programming (standard minimization problem)

$$\min_{\mathbf{y}\in\mathbb{R}^m} \mathbf{y}^T\mathbf{b},$$

subject to the constraints

$$\mathbf{y}^T\mathbf{A} \geq \mathbf{c}^T,$$

and $y_1 \geq 0, \cdots, y_m \geq 0$. Here $\mathbf{A}$ is $m \times n$ matrix, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$.

The dual maximization problem

$$\max_{\mathbf{x}\in\mathbb{R}^n} \mathbf{c}^T\mathbf{x}, \quad s.t.\ \mathbf{A}\mathbf{x} \leq \mathbf{b} \text{ and } \mathbf{x} \geq 0.$$

Note that the linear quantile regression model can be rewritten as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + e_i = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + (u_i - v_i),$$

where $u_i = e_i I(e_i > 0)$, $v_i = |e_i| I(e_i < 0)$.

$$\begin{aligned}
& \text{Therefore, } \min_{\mathbf{b}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \mathbf{b}) \\
\Leftrightarrow \quad & \min_{\{\mathbf{b}, \boldsymbol{u}, \boldsymbol{v}\}} \tau 1_n^T \boldsymbol{u} + (1-\tau) 1_n^T \boldsymbol{v} \\
& s.t. \;\; \mathbf{y} - \mathbf{X}^T \mathbf{b} = \boldsymbol{u} - \boldsymbol{v} \\
& \mathbf{b} \in \mathbb{R}^p, \;\; \boldsymbol{u} \geq 0, \;\; \boldsymbol{v} \geq 0.
\end{aligned}$$

This is a standard linear programming (minimization) program.

2.2 Basic Properties

1. **Basic equivariance properties**. Let A be any $p \times p$ nonsingular matrix, $\boldsymbol{\gamma} \in \mathbb{R}^p$, and $a > 0$ is a constant. Let $\hat{\boldsymbol{\beta}}(\tau; y, \mathbf{X})$ be the estimator in the τth quantile regression based on observations $(y, \mathbf{X})$. Then for any $\tau \in [0, 1]$,

   (i) $\hat{\boldsymbol{\beta}}(\tau; ay, \mathbf{X}) = a\hat{\boldsymbol{\beta}}(\tau; y, \mathbf{X})$;

   (ii) $\hat{\boldsymbol{\beta}}(\tau; -ay, \mathbf{X}) = -a\hat{\boldsymbol{\beta}}(1 - \tau; y, \mathbf{X})$;

   (iii) $\hat{\boldsymbol{\beta}}(\tau; y + \mathbf{X}\gamma, \mathbf{X}) = \hat{\boldsymbol{\beta}}(\tau; y, \mathbf{X}) + \boldsymbol{\gamma}$;

   (iv) $\hat{\boldsymbol{\beta}}(\tau; y, \mathbf{X}A) = A^{-1}\hat{\boldsymbol{\beta}}(\tau; y, \mathbf{X})$.

2. **Equivariance property:** quantiles are equivariant to monotone transformations. Suppose $h(\cdot)$ is an increasing function on $\mathbb{R}$. Then for any variable Y,

$$Q_{h(Y)}(\tau) = h\left\{Q_\tau(Y)\right\}.$$

3. **Interpolation**: a basic solution from LP interpolates p observations. If the first column of the design matrix is one corresponding to the intercept, then there are at least p zero, and at most $n\tau$ negative and $n(1-\tau)$ positive residuals $y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau)$.

2.3 Subgradient Condition

Define

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}).$$

- Piecewise linear and continuous.
- Differentiable except at points such that $y_i - \mathbf{x}_i^T \boldsymbol{\beta} = 0$.

The **directional derivative** of $R(\boldsymbol{\beta})$ in the direction $\boldsymbol{w}$

$$\nabla R(\boldsymbol{\beta}, \boldsymbol{w}) \quad = \quad \frac{d}{dt} R(\boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{w} t)|_{t=0}.$$

Note that

$$
\begin{aligned}
& \frac{d}{dt}\rho_\tau(y - \mathbf{x}^T\boldsymbol{\beta} - \mathbf{x}^T\boldsymbol{w}t)|_{t=0} \\
= \quad & \frac{d}{dt}(y - \mathbf{x}^T\boldsymbol{\beta} - \mathbf{x}^T\boldsymbol{w}t)\left\{\tau - I(y - \mathbf{x}^T\boldsymbol{\beta} - \mathbf{x}^T\boldsymbol{w}t < 0)\right\}|_{t=0} \\
= \quad & \begin{cases} -\mathbf{x}^T\boldsymbol{w}\tau, & y - \mathbf{x}^T\boldsymbol{\beta} > 0 \\ -\mathbf{x}^T\boldsymbol{w}(1-\tau), & y - \mathbf{x}^T\boldsymbol{\beta} < 0 \\ -\mathbf{x}^T\boldsymbol{w}\{\tau - I(-\mathbf{x}^T\boldsymbol{w} < 0)\}, & y - \mathbf{x}^T\boldsymbol{\beta} = 0 \end{cases} \\
= \quad & \mathbf{x}^T\boldsymbol{w}\psi_\tau^*(y - \mathbf{x}^T\boldsymbol{\beta}, -\mathbf{x}^T\boldsymbol{w}),
\end{aligned}
\tag{2.1}
$$

where

$$
\psi_\tau^*(u, v) = \begin{cases} \tau - I(u < 0), & u \neq 0 \\ \tau - I(v < 0), & u = 0. \end{cases}
$$

Thus

$$\nabla R(\boldsymbol{\beta}, \boldsymbol{w}) \quad = \quad \sum_{i=1}^{n} \mathbf{x}_i^T \boldsymbol{w} \psi_\tau^*(y_i - \mathbf{x}_i^T \boldsymbol{\beta}, -\mathbf{x}_i^T \boldsymbol{w}). \qquad (2.2)$$

Note

$$\begin{aligned} & \nabla R(\hat{\boldsymbol{\beta}}, \boldsymbol{w}) \geq 0 \text{ for all } \boldsymbol{w} \in \mathbb{R}^p \text{ with } \|\boldsymbol{w}\| = 1 \\ \Leftrightarrow \quad & \hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} R(\boldsymbol{\beta}). \end{aligned}$$

Theorem 1 *If (y, X) are in general positions (i.e. if any p observations of them yield a unique exact fit), then there exists a minimizer of $R(\boldsymbol{\beta})$ of the form $b(h) = X(h)^{-1}y(h)$ if and only if, for some $h \in \mathcal{H}$,*

$$-\tau 1_p \leq \boldsymbol{\xi}(h) \leq (1-\tau)1_p,$$

where $\boldsymbol{\xi}(h)^T = \sum_{i \in \bar{h}} \psi_\tau\{y_i - \mathbf{x}_i^T \mathbf{b}(h)\}\mathbf{x}_i^T X(h)^{-1}$, and $\bar{h}$ is the complement of h.

Proof. In linear programming, vertex solutions (**basic solutions**) correspond to points at which p observations are interpolated, i.e. $\big(y(h), X(h)\big) = \{(y_i, \mathbf{x}_i), i \in h\}$. That is, the basic solutions pass through these n points as

$$\mathbf{b}(h) = X(h)^{-1}y(h), \quad h \in \mathcal{H}^* = \{h \in \mathcal{H}^* : |X(h)| \neq 0\}.$$

For any $\boldsymbol{w} \in \mathbb{R}^p$, reparameterize to get $\boldsymbol{v} = X(h)\boldsymbol{w}$, i.e. $\boldsymbol{w} = X(h)^{-1}\boldsymbol{v}$.

For a basic solution $\mathbf{b}(h)$ to be the minimizer, we need for all $\boldsymbol{v} \in \mathbb{R}^p$,

$$-\sum_{i=1}^{n} \psi_\tau^*\{y_i - \mathbf{x}_i^T\mathbf{b}(h), -\mathbf{x}_i^T X(h)^{-1}\boldsymbol{v}\}\mathbf{x}_i^T X(h)^{-1}\boldsymbol{v} \geq 0. \tag{2.3}$$

WLOG, assume $X(h) = (\mathbf{x}_1^T, \cdots, \mathbf{x}_p^T)^T$.

- If $i \in h$, $\mathbf{x}_i^T X(h) = \mathbf{e}_i^T$, where $\mathbf{e}_i$ is a p-dimensional vector containing all zeros except the ith element being of 1. Thus $\mathbf{e}_i\boldsymbol{v} = v_i$.

- If (y, X) are in general position, none of the residuals $y_i - \mathbf{x}_i^T \mathbf{b}(h)$ with $i \in \bar{h}$ is zero. If y_i's have a density wrt Lesbesgue measure, then with probability one (y, X) are in general position.

- The space of directions $\boldsymbol{v} \in \mathbb{R}^p$ is spanned by $\boldsymbol{v} = \pm \mathbf{e}_k, k = 1, \cdots, p$. So (2.3) holds for any $\boldsymbol{v} \in \mathbb{R}^p$ iff the inequality holds for $\pm \mathbf{e}_k, k = 1, \cdots, p$.

Therefore, (2.3) becomes

$$0 \leq -\sum_{i \in h} \psi_\tau^* \{0, -v_i\} v_i - \boldsymbol{\xi}(h)^T \boldsymbol{v}, \tag{2.4}$$

where $\boldsymbol{\xi}(h)^T = \sum_{i \in \bar{h}} \psi_\tau \{y_i - \mathbf{x}_i^T \mathbf{b}(h)\} \mathbf{x}_i^T X(h)^{-1}$.

- If $\boldsymbol{v} = \mathbf{e}_i$, we have

$$0 \leq -(\tau - 1) - \xi_i(h), \quad i = 1, \cdots, p.$$

- If $\boldsymbol{v} = -\mathbf{e}_i$, we have

$$0 \leq \tau + \xi_i(h), \quad i = 1, \cdots, p.$$

That is,

$$-\tau 1_p \leq \boldsymbol{\xi}(h) \leq (1-\tau) 1_p.$$

$\square$

Remark 1 *The total score:*

$$\Big\| \sum_{i=1}^{n} \mathbf{x}_i \psi_\tau\{y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)\} \Big\| \leq Cp \max_{i=1,\cdots,n} \|\mathbf{x}_i\|.$$

2.4 Consistency

Coefficient estimator in linear quantile regression model

$$\hat{\boldsymbol{\beta}}(\tau) = argmin_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\mathbf{b}).$$

Classical Sufficient Regularity Conditions

A1 The distribution functions of Y given $\mathbf{x}_i$, $F_i(\cdot)$, are absolutely continuous with continuous densities $f_i(\cdot)$ that are uniformly bounded away from 0 and ∞ at $\xi_i(\tau) = Q_\tau(Y|\mathbf{x}_i)$.

A2 There exist positive definite matrices D_0 and D_1 such that

(i) $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T = D_0$;

(ii) $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n f_i\{\xi_i(\tau)\}\mathbf{x}_i\mathbf{x}_i^T = D_1$;

(iii) $\max_{i=1,\cdots,n} \|\mathbf{x}_i\| = o(n^{1/2})$.

Theorem 2 *Under Conditions A1 and A2 (i),* $\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{p} \boldsymbol{\beta}(\tau)$.

Sketch of the proof:

1. Define $\bar{\rho}_\tau(y - \mathbf{x}^T\mathbf{b}) = \rho_\tau(y - \mathbf{x}^T\mathbf{b}) - \rho_\tau\{y - \mathbf{x}^T\boldsymbol{\beta}(\tau)\}$.
2. Use the uniform law of large numbers to show that

$$\sup_{\mathbf{b}\in\mathcal{B}} n^{-1}\sum_{i=1}^{n}\left[\bar{\rho}_\tau(y_i - \mathbf{x}_i^T\mathbf{b}) - E\left\{\bar{\rho}_\tau(y_i - \mathbf{x}_i^T\mathbf{b})\right\}\right] = o_p(1),$$

   where $\mathcal{B}$ is a compact subset of $\mathbb{R}^p$. **Reference:** Pollard (1991).
3. Note that $\hat{\boldsymbol{\beta}}(\tau) \to \boldsymbol{\beta}(\tau)$ holds if for any $\epsilon > 0$, $\bar{Q}(\mathbf{b}) \equiv n^{-1}\sum_{i=1}^{n} E\left\{\bar{\rho}_\tau(y_i - \mathbf{x}_i^T\mathbf{b})\right\}$ is bounded away from zero with probability approaching one for any $\|\mathbf{b} - \boldsymbol{\beta}(\tau)\| \geq \epsilon$; see e.g. Lemma 2.2 of White (1980).
4. Under Conditions A1 and A2 (i), $\bar{Q}(\mathbf{b})$ has a unique minimizer $\boldsymbol{\beta}(\tau)$ and Step 3 goes through.
5. The convergence is thus proven.

2.5 Asymptotic Normality

Theorem 3 *Under Conditions A1 and A2,*

$$n^{1/2}\left\{\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)\right\} \xrightarrow{d} N\left(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}\right).$$

For the $i.i.d.$ error models, i.e. $f_i\{\xi_i(\tau)\} = f_\epsilon(0)$, the above result can be simplified as

$$n^{1/2}\left\{\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)\right\} \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f_\epsilon^2(0)}D_0^{-1}\right).$$

$D_0 \approx n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, but D_1 is harder to compute.

Sketch of the proof

1. The solution satisfies $n^{-1/2}\sum_{i=1}^n \mathbf{x}_i\psi_\tau\{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau)\} = o_p(1)$.
2. $n^{-1}\sum_{i=1}^n \mathbf{x}_i\big(\psi_\tau\{y_i - \mathbf{x}_i^T\boldsymbol{\beta}\} - \psi_\tau\{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0\}\big)$ can be well approximated by it expectation $b(\boldsymbol{\beta}) - b(\boldsymbol{\beta}_0)$, uniformly for $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$.
3. Plug in the estimate $\hat{\boldsymbol{\beta}}(\tau)$, and then use Taylor expansion on $b(\boldsymbol{\beta}) - b(\boldsymbol{\beta}_0) = D_1(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \cdots$.
4. $n^{-1/2}\sum_{i=1}^n \mathbf{x}_i\psi_\tau\{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau)\} = -D_1(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0) + \cdots$

   Reference: He and Shao (1996).

3 Inference

3.1 Wald-type Test

3.1.1 Asymptotic Normality

- Asymptotic normality in $i.i.d.$ settings

$$n^{1/2}\left\{\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)\right\}\xrightarrow{d}N\left(0,\frac{\tau(1-\tau)}{f_\epsilon^2(0)}D_0^{-1}\right),$$

where $D_0=\lim_{n\to\infty}n^{-1}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T$.

- Asymptotic normality in non-$i.i.d.$ settings

$$n^{1/2}\{\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)\} \xrightarrow{d} N\left(0, \tau(1-\tau)D_1(\tau)^{-1}D_0D_1^{-1}(\tau)\right),$$

where

$$D_1(\tau)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} f_i\{\mathbf{x}_i^T\boldsymbol{\beta}(\tau)\}\mathbf{x}_i\mathbf{x}_i^T.$$

- Asymptotic covariance between quantiles

$$\begin{aligned} & \mathsf{Acov}\Big(\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau_i)-\boldsymbol{\beta}(\tau_i)\}, \sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau_j)-\boldsymbol{\beta}(\tau_j)\}\Big) \\ = \quad & (\tau_i \wedge \tau_j - \tau_i\tau_j)D_1(\tau_i)^{-1}D_0D_1^{-1}(\tau_j). \end{aligned}$$

3.1.2 Wald Test for General Linear Hypotheses

Define the coefficient vector $\boldsymbol{\theta} = (\boldsymbol{\beta}(\tau_1)^T, \ldots, \boldsymbol{\beta}(\tau_m)^T)^T$.

- Null hypothesis $H_0 : R\boldsymbol{\theta} = \boldsymbol{r}$.
- Test statistic

$$T_n = n(R\hat{\boldsymbol{\theta}} - \boldsymbol{r})^T (RV^{-1}R^T)^{-1}(R\hat{\boldsymbol{\theta}} - \boldsymbol{r}),$$

  where V is the $mp \times mp$ matrix with the ijth block

$$V(\tau_i, \tau_j) = (\tau_i \wedge \tau_j - \tau_i \tau_j) D_1(\tau_i)^{-1} D_0 D_1^{-1}(\tau_j).$$

- Under H_0, $T_n \xrightarrow{d} \chi^2_q$, where q is the rank of R.
- **Drawback**: the covariance matrix involves the unknown density functions (nuisance parameters), i.e. $f_i\{\mathbf{x}_i^T \boldsymbol{\beta}(\tau)\}$ in Non-IID settings, and $f_\epsilon(0)$ in IID settings.
- Reference: Koenker and Machado (1999).

3.1.3 Estimation of Asymptotic Covariance Matrix

1. IID settings:

$$\mathrm{var}\{n^{1/2}\hat{\boldsymbol{\beta}}(\tau)\} \approx \frac{\tau(1-\tau)}{\hat{f}_\epsilon^2(0)}\hat{D}_0^{-1},$$

where $\hat{D}_0 = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$.

Estimation of $f_\epsilon(0) = f_\epsilon\{F_\epsilon^{-1}(\tau)\}$

- Sparsity parameter:

$$s(\tau) = \frac{1}{f\{F^{-1}(\tau)\}}.$$

- Note $F\{F^{-1}(t)\} = t$. Differentiate both side with respect to t, we get

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

That is, the sparsity parameter $s(t)$ is simply the derivative of quantile function $F^{-1}(t)$ wrt t.

- **Difference quotient estimator**

$$\hat{s}_n(t) = \frac{\hat{F}^{-1}(t + h_n|\bar{\mathbf{x}}) - \hat{F}^{-1}(t - h_n|\bar{\mathbf{x}})}{2h_n},$$

where $h_n \to 0$ as $n \to \infty$, and $\hat{F}^{-1}(t|\bar{\mathbf{x}}) = \bar{\mathbf{x}}^T \hat{\boldsymbol{\beta}}(t)$ is the estimated tth conditional quantile of Y given $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$. This is advantageous in small to moderate sample sizes where computing the whole process $\hat{\boldsymbol{\beta}}(\tau)$ is tractable. For large samples, it is preferable to use residual-based estimator

$$\hat{s}_n(t) = \frac{\hat{F}_n^{-1}(t + h_n) - \hat{F}_n^{-1}(t - h_n)}{2h_n},$$

where $\hat{F}_n^{-1}(\cdot)$ is the empirical quantile function of estimated residuals $\hat{\epsilon}_i = y_i - \mathbf{x}_i^T \hat{\beta}(\tau)$, $i = 1, \ldots, n$.

Non-IID settings:

$$\mathrm{var}\{n^{1/2}\hat{\boldsymbol{\beta}}(\tau)\} \approx \tau(1-\tau)\hat{D}_1(\tau)^{-1}\hat{D}_0\hat{D}_1^{-1}(\tau).$$

The main challenge is the estimation of $D_1(\tau)$.

- **Hendricks-Koenker Sandwich**
  - Suppose the conditional quantiles of Y given $\mathbf{x}$ are linear at quantile levels around τ.
  - Then fit quantile regression at $(\tau \pm h_n)$th quantiles, resulting in $\hat{\boldsymbol{\beta}}(\tau - h_n)$ and $\hat{\boldsymbol{\beta}}(\tau + h_n)$.
  - Estimate $f_i\{\xi_i(\tau)\}$ by

$$\tilde{f}_i\{\xi_i(\tau)\} = \frac{2h_n}{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau + h_n) - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau - h_n)},$$

    where $\xi_i(\tau) = Q_\tau(Y|\mathbf{x}_i)$.
  - In finite sample studies, quantiles may cross so that the upper quantiles may be estimated to be smaller than lower quantiles.

A modified estimator to account for this issue:

$$\hat{f}_i\{\xi_i(\tau)\} = \max\left(0, \frac{2h_n}{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau + h_n) - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\tau - h_n) - \epsilon}\right),$$

where ϵ is a small positive constant to avoid zero denominator.

– Estimator of $D_1(\tau)$:

$$\hat{D}_1(\tau) = n^{-1}\sum_{i=1}^{n}\hat{f}_i\{\xi_i(\tau)\}\mathbf{x}_i\mathbf{x}_i^T.$$

3.2 Rank Score Test

Consider the model

$$Q_\tau(Y|\mathbf{x}_i, \boldsymbol{z}_i) = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + \boldsymbol{z}_i^T \boldsymbol{\gamma}(\tau),$$

and hypotheses

$$H_0 : \boldsymbol{\gamma}(\tau) = 0 \;\; \text{v.s.} \;\; H_a : \boldsymbol{\gamma}(\tau) \neq 0.$$

Here $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$ and $\boldsymbol{\gamma}(\tau) \in \mathbb{R}^q$.

Score function:

$$S_n = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{z}_i^* \psi_\tau \{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\tau)\},$$

- $\psi_\tau(u) = \tau - I(u < 0)$;
- $\boldsymbol{Z}^* = (\boldsymbol{z}_i^*) = \boldsymbol{Z} - \mathbf{X}(\mathbf{X}^T \Psi \mathbf{X})^{-1} \mathbf{X}^T \Psi \boldsymbol{Z}$,
- $\Psi = \text{diag}\left(f_i\{Q_\tau(Y|\mathbf{x}_i, \boldsymbol{z}_i)\}\right)$;
- $\hat{\boldsymbol{\beta}}(\tau)$ is the quantile coefficient estimator obtained under H_0.

Asymptotic property: Under H_0, as $n \to \infty$,

$$S_n = AN(0, M_n^{1/2}), \tag{3.1}$$

where $M_n = n^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i^* \boldsymbol{z}_i^{*T} \tau(1-\tau)$.

Rank-score test statistic:

$$T_n = S_n^T M_n^{-1} S_n \xrightarrow{d} \chi_q^2, \quad \text{under } H_0.$$

Simplification for *i.i.d.* settings

- $\boldsymbol{Z}^* = (\boldsymbol{z}_i^*) = \left\{I - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right\}\boldsymbol{Z}$: the residuals by projecting $\boldsymbol{Z}$ on $\mathbf{X}$;
- $M_n = \tau(1-\tau)n^{-1}\sum_{i=1}^n \boldsymbol{z}_i^* \boldsymbol{z}_i^{*T}$;
- so no need to estimate the nuisance parameters $f_i\{Q_\tau(Y|\mathbf{x}_i, \boldsymbol{z}_i)\}$.

Construction of confidence interval (CI) of $\gamma(\tau)$

- $\gamma(\tau)$ is a parameter of interest, corresponding to one of the covariates.
- CI of $\gamma(\tau)$ can be constructed by inversion of rank score test.
- Consider the hypotheses

$$H_0 : \gamma(\tau) = \gamma_0 \;\; \text{v.s.} \;\; H_a : \gamma(\tau) \neq \gamma_0,$$

  where γ_0 is a prespecified scalar.
- Reject H_0 if $T_n \geq \chi^2_\alpha(1)$, the $(1-\alpha)$th quantile of $\chi^2(1)$, and vice versa.
- The collection of all the $\boldsymbol{\gamma}_0$ for which H_0 is not rejected is taken to be the $(1-\alpha)$th CI of $\gamma(\tau)$.
- Reference: Koenker (2005)

A special case for illustration

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_i + e_i$$

Hypothesis H_0: $\beta_1(\tau) = 0$.

The quantile rank score test is used on

$$S_n = n^{-1/2} \sum_i (x_i - \bar{x})\psi_\tau(y_i - Q(\tau))$$

where $Q(\tau)$ is the τ-th quantile of $\{y_i\}$.

c.f. Sign test at $\tau = 0.5$.

3.3 Bootstrap

Idea of the bootstrap

- Data $X_1, \cdots, X_n$ from F_θ.
- We can estimate θ from $T(F_n)$, where F_n is the empirical distribution of the sample.
- If we know F, we can draw samples of size n from F, and get many copies of $\hat{\theta}$ to obtain the variance of the estimate.
- The bootstrap uses F_n as an approximation to F, and draws samples from F_n instead.

3.3.1 Residual Bootstrap

For $i.i.d.$ errors, location-shift model $y_i = \mathbf{x}_i^T \beta(\tau) + \epsilon_i$:

- Obtain the estimator $\hat{\beta}(\tau)$ using the observed sample, and residuals $\hat{\epsilon}_i = y_i - \mathbf{x}_i^T \hat{\beta}(\tau)$.
- Draw bootstrap samples $\epsilon_i^*, i = 1, \cdots, n$ from $\{\hat{\epsilon}_i, i = 1, \cdots, n\}$ with replacement, and define $y_i^* = \mathbf{x}_i^T \hat{\beta}(\tau) + \epsilon_i^*$.
- Compute the bootstrap estimator $\hat{\beta}^*(\tau)$ by quantile regression using the bootstrap sample.
- Carry out inference by calculating the covariance of $\hat{\boldsymbol{\beta}}(\tau)$ by the sample covariance of bootstrap estimators or construct CI using percentile methods.

3.3.2 Paired Bootstrap

- Generate bootstrap sample $(y_i^*, \mathbf{x}_i^*)$ by drawing with replacement from the n pairs $\{(y_i, \mathbf{x}_i), i = 1, \cdots, n\}$.
- Obtain the bootstrap estimator $\hat{\beta}^*(\tau)$ by quantile regression using the bootstrap sample.

3.3.3 MCMB

Markov chain marginal bootstrap (mcmb) (He and Hu, 2002, Kocherginsky, Mu and He, 2005). Instead of solving a p-dimensional estimating equation for each bootstrap replication, MCMB solves p one-dimensional estimating equations.

Model:

$$Q_\tau(Y_i|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}(\tau), \quad \boldsymbol{\beta}(\tau) \in \mathbb{R}^p,$$

where $\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,p})^T$.

Procedure

(i) Calculate $r_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\tau)$. Define $\boldsymbol{z}_i = \mathbf{x}_i \psi_\tau(r_i) - \bar{\boldsymbol{z}}$ with $\bar{\boldsymbol{z}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \psi_\tau(r_i)$, where $\psi_\tau(r) = \tau - I(r < 0)$.

(ii) Step 0: let $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}(\tau)$.

(iii) Step k: for each integer $1 \le j \le p$ in the ascending order, draw with replacement from $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ to obtain $z_1^{j,k}, \cdots, z_n^{j,k}$. Solve $\beta_j^{(k)}$ as the solution to

$$\sum_{i=1}^n x_{i,j} \psi_\tau \left\{ y_i - \sum_{l<j} x_{i,l} \beta_l^{(k)} - \sum_{l>j} x_{i,l} \beta_l^{(k-1)} - x_{i,j} \beta_k^{(k)} \right\} = \sum_{i=1}^n z_i^{j,k}.$$

(iv) Repeat Step (iii) until K replications $\boldsymbol{\beta}^{(k)}, k = 1, \cdots, K$ are obtained. The variance of $\hat{\boldsymbol{\beta}}(\tau)$ is then estimated by the sample variance of $\{\boldsymbol{\beta}^{(k)}, k = 1, \cdots, K\}$.

Some Other Bootstrap Methods

- Bootstrap estimating equations: Parzen, Ying, and Wei (1994).
- Generalized bootstrap: Bose and Chatterjee (2003).
- Wild bootstrap: Feng, He and Hu (2011).
- Bayesian methods: Yang, Wang and He (2016).

Recommendations:

- Rank-score methods are quite reliable unless some covariate is heavily skewed.
- Paired bootstrap is slightly conservative.
- Wald-type methods are all right for large-sample problems.
- MCMB is useful when the dimension is high.

4 Algorithms and Computer Code

given by Professor Ying Wei

5 Examples

given by Professor Ying Wei

6 Nonparametric quantile curves

6.1 Introduction

Given data (x_i, y_i), want to capture the dependence of y on x.
Regression model:

$$y_i = f(x_i) + e_i.$$

Nonparametric Models

- Motivation: the underlying regression function is so complicated that no reasonable parametric model would be adequate
- Do not assume any specific form of f. More flexible.
- Infinite dimensional parameters.

6.2 Local Polynomial

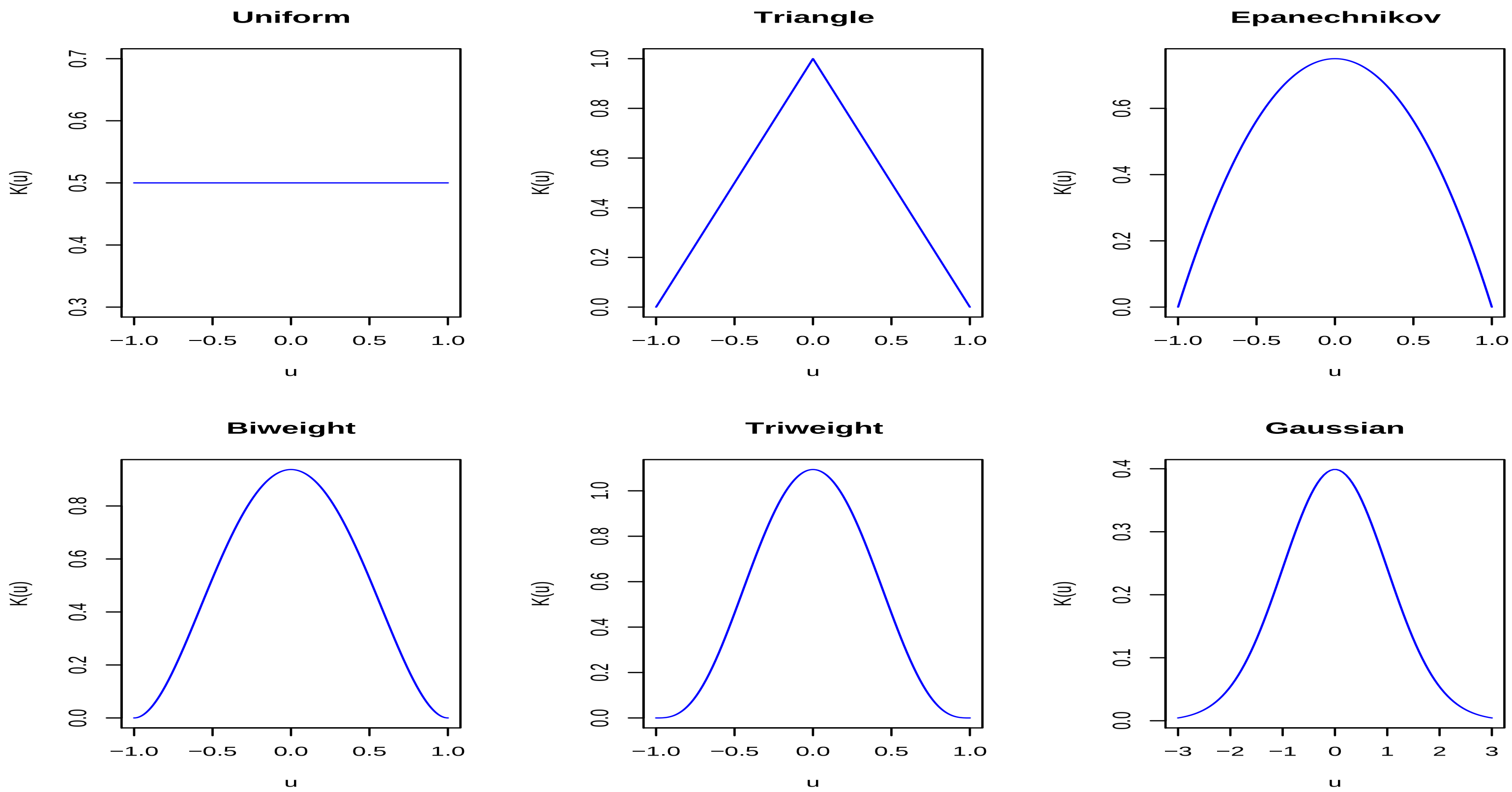
Local constant quantile regression

Define $f_\tau(x) = Q_\tau(Y|x = x)$: τth conditional quantile of Y given $X = x$. That is $f_\tau(x) = argmin_a E\{\rho_\tau(Y - a)|X = x\}$.
The local constant quantile estimator of $f_\tau(x)$ is

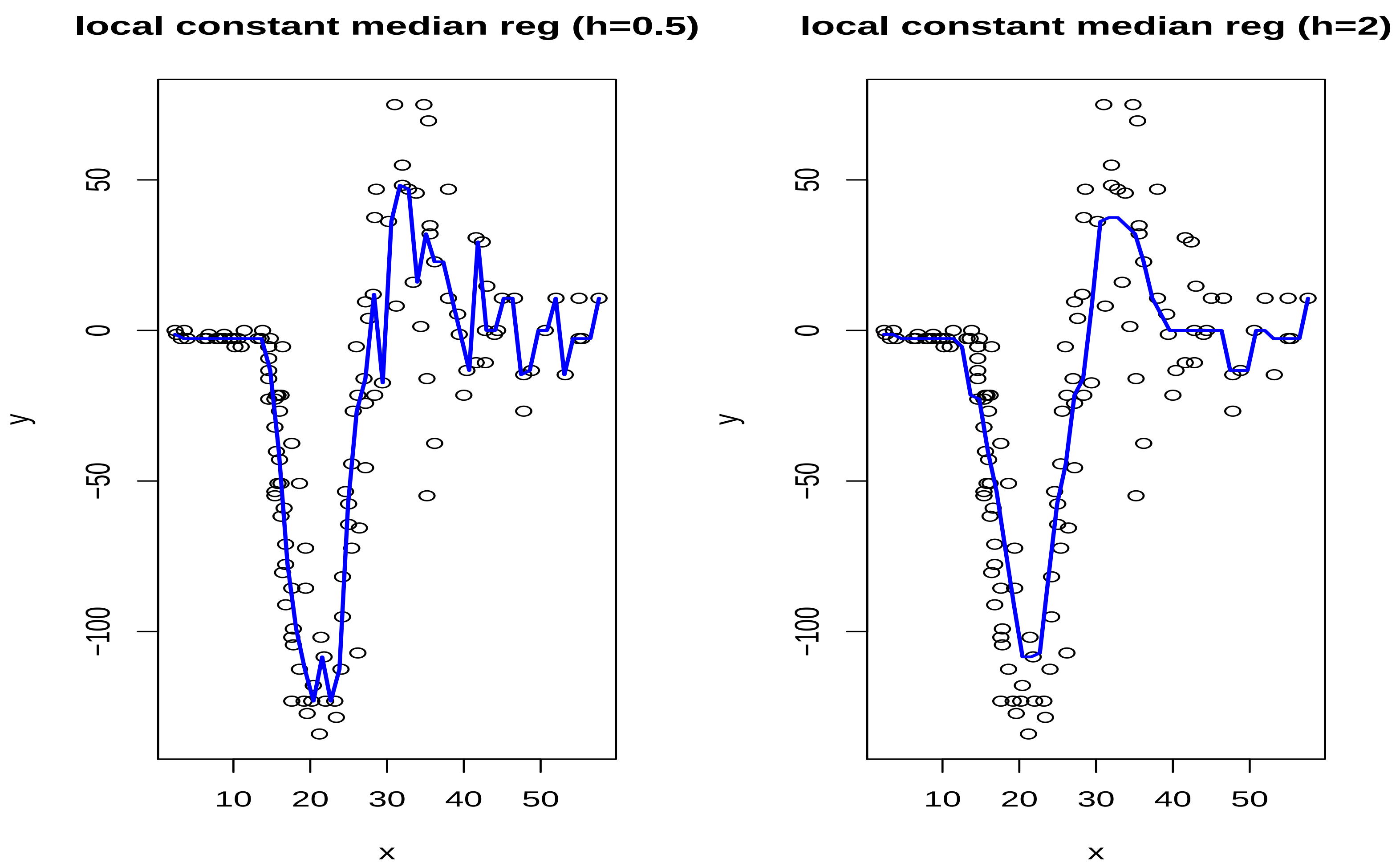
$$\hat{f}_\tau(x) = argmin_a \sum_{i=1}^{n} \rho_\tau(y_i - a) K\left(\frac{x - x_i}{h}\right),$$

- $h > 0$ is the bandwidth parameter,
- $K(\cdot)$ is the kernel function,
- points within $[x - h, x + h]$ receive positive weights (except Gaussian kernel).

Kernel functions



Local constant rq



Local linear quantreg

- Approximate $f_\tau(x)$ by a linear function

$$f_\tau(z) = f_\tau(x) + f_\tau'(x)(z-x) \doteq a + b(z-x),$$

for z in a neighborhood of x.

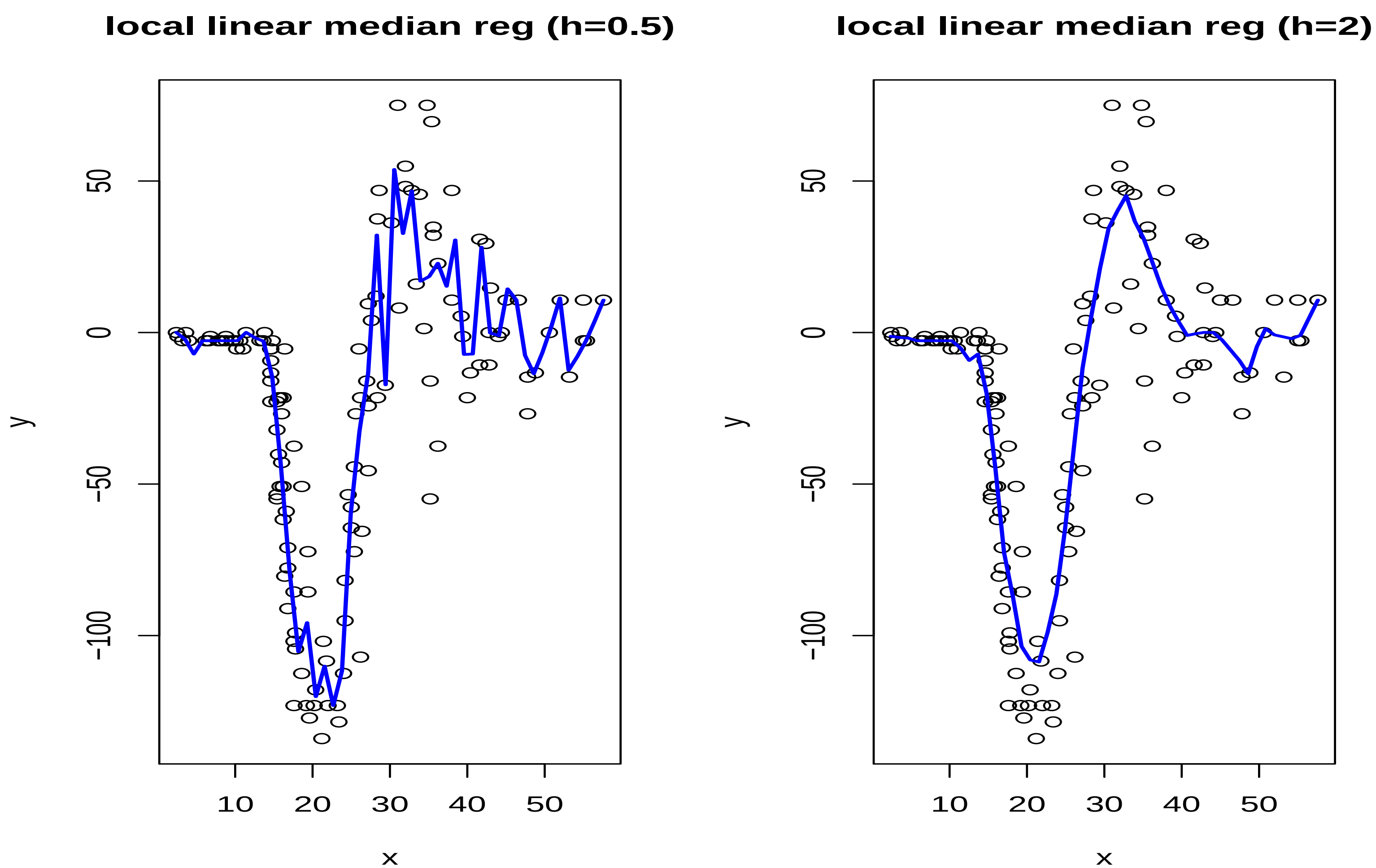
- Estimating $f_\tau(x)$ is equivalent to estimating a

- Estimating $f_\tau'(x)$ is equivalent to estimating b

- Local linear estimator of $f_\tau(x)$ is $\hat{f}_\tau(x) = \hat{a}$, where $\hat{a}$ and $\hat{b}$ minimize

$$\sum_{i=1}^{n} \rho_\tau \left\{ y_i - a - b(x_i - x) \right\} K\left(\frac{x - x_i}{h} \right).$$

Local linear quantreg



How to choose h

- When estimating $f(x)$, only points within $[x-h, x+h]$ receive positive weights (except Gaussian kernel).
- smaller h: rougher estimates, relying heavily on the data near x, smaller bias, larger variance
- larger h: more averaging range, smoother estimates, larger bias, smaller variance

Bandwidth Selection (m-fold cross validation)

- Randomly divide the data into m non-overlapped and roughly equal-sized parts $D_1, \cdots, D_m$.

- For the ith part, fit the model using the data from the test data, "predict" the τth conditional quantiles, and calculate the quantile prediction error as

$$\sum_{j \in D_i} \rho_\tau \left\{ Y_j - \hat{f}_\tau(x_j)_{-D_i} \right\}.$$

- Repeat this procedure for $i = 1, \cdots, m$, and calculate the averaged quantile prediction error.

- Select h with the smallest averaged prediction error.

6.3 B-splines

- B-splines are piecewise polynomials that are smoothly connected at the knots.

- B-spline representation is via a series of polynomial basis functions which have local support.

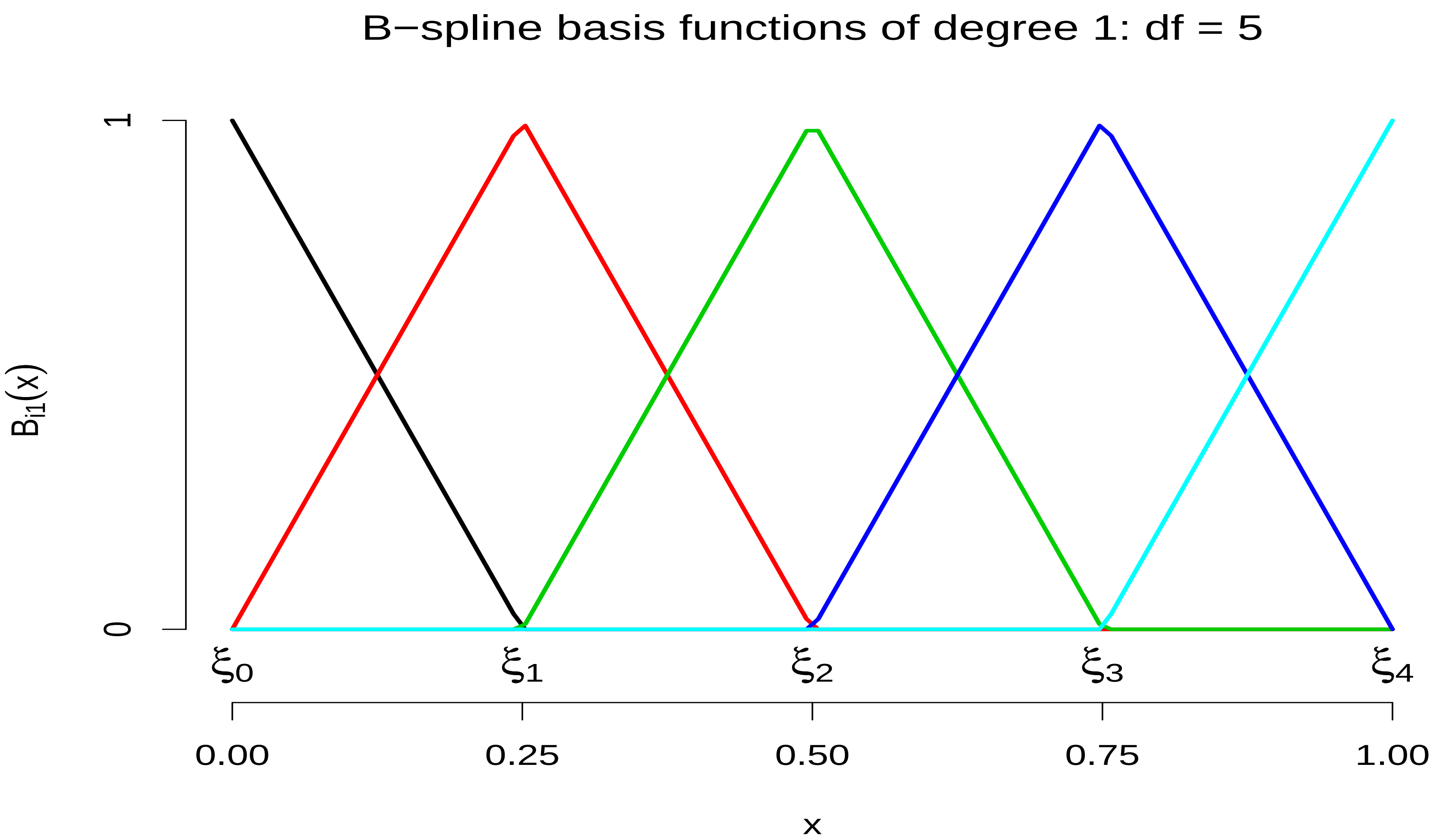
- Consider $x \in [0, 1]$ with $K = 3$ internal knots

$$\xi = (0.25, 0.50, 0.75)$$

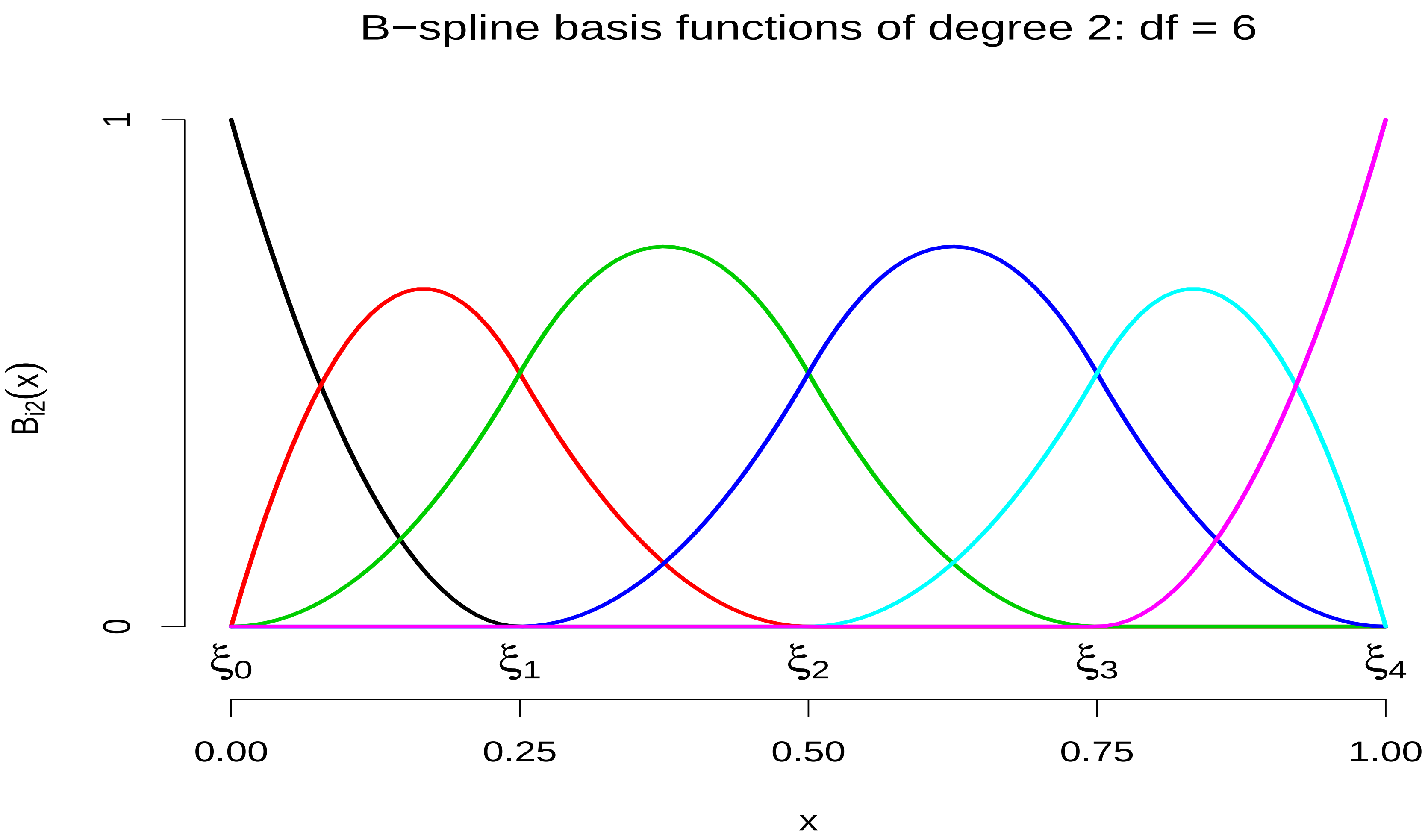
- Include the boundary knots:

$$\xi = (0, 0.25, 0.50, 0.75, 1), \quad \xi_0 = 0, \cdots, \xi_{K+1} = \xi_4 = 1.$$

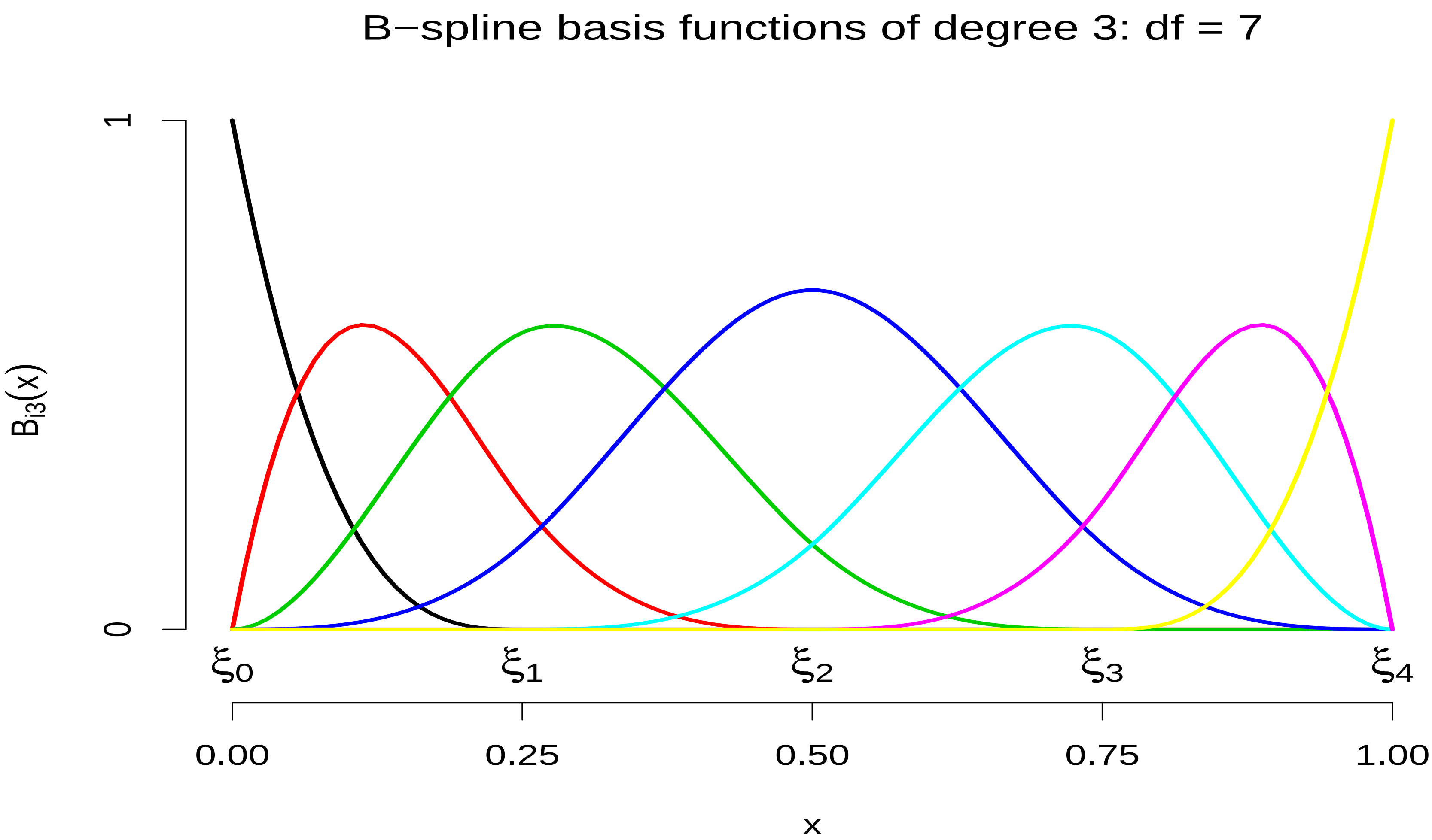
B-spline basis functions of degree 1: df=5



B-spline basis functions of degree 2: df=6



B-spline basis functions of degree 3: df=7



Summary: degree-p B-spline

- Define an augmented knot sequence:

$$\xi = (\xi_{-p}, \cdots, \xi_0, \xi_1, \cdots, \xi_K, \xi_{K+1}, \cdots, \xi_{K+p+1})$$

- For $i = -p, \cdots, K+p$, let

$$B_{i,0}(x) = \begin{cases} 1 & x \in [\xi_i, \xi_{i+1}) \\ 0 & \text{otherwise} \end{cases},$$

where $B_{i,0}(x) = 0$ if $\xi_i = \xi_{i+1}$.

- The ith B-spline basis function of degree j, $j = 1, \cdots, p$ is given by

$$B_{i,j}(x) = \frac{x - \xi_i}{\xi_{i+j} - \xi_i} B_{i,j-1}(x) + \frac{\xi_{i+j+1} - x}{\xi_{i+j+1} - \xi_{i+1}} B_{i+1,j-1}(x),$$

for $i = -p, \cdots, 0, \cdots, K+p-j$.

How to estimate the quantile function given knots?

Given the B-spline basis functions of order p, the normalized basis functions add up to one, and the vector of basis functions is denoted by $\pi(x)$.

We approximate

$$f_\tau(x) = \pi(x)^T \alpha$$

for some coefficient α, and then estimate it by

$$\hat{\alpha} = argmin_\alpha \sum_i \rho_\tau(y_i - \pi(x_i)^T \alpha),$$

and

$$\hat{f}_\tau(x) = \pi(x)^T \hat{\alpha}.$$

How to choose K and knot locations

- For B-splines of degree p, suppose there are K internal knots, the knot locations can be chosen as the $i/(K+1)$th sample quantiles of x, $i = 1, \cdots, K$.

- The number of knots K can be chosen by minimizing the Schwartz Information Criterion

$$SIC(K) = \log\left[\sum_{i=1}^{n} \rho_\tau\{y_i - \hat{f}_\tau(x_i)\}\right] + \frac{\log n}{n} edf,$$

where $edf = K + p + 1$ is the number of parameters in the model.

6.4 Quantile smoothing splines

- Estimate $f(\cdot)$ via minimizing the penalized objective function:

$$RSS(f, \tau, \lambda) = \sum_{i=1}^{n} \rho_\tau\{y_i - f(x_i)\} + \lambda V(f'),$$

  - $V(f') = \sum_{i=1}^{n-1} |f'(x_{i+1}) - f'(x_i)|$ is the total variation penalty on f'
  - λ is the smoothing parameter

- Basic property: the function f minimizing $RSS(f, \tau, \lambda)$ is a linear spline with knots at the points $x_1, \cdots, x_n$.

- The solution at a general $\tau \in (0, 1)$ can be obtained by using linear programming.

- Reference: Koenker, Ng, and Portnoy (1994)

6.5 Extensions

- additive model with $f_\tau(x_1, x_2) = f_1(x_1) + f_2(x_2)$
- partially linear model with $f_\tau(x, z) = x^T\beta_\tau + g_\tau(z)$
- single-index models with $f_\tau(x) = g_\tau(x^T\beta_\tau)$

7 Censored quantile regression

7.1 Background

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

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

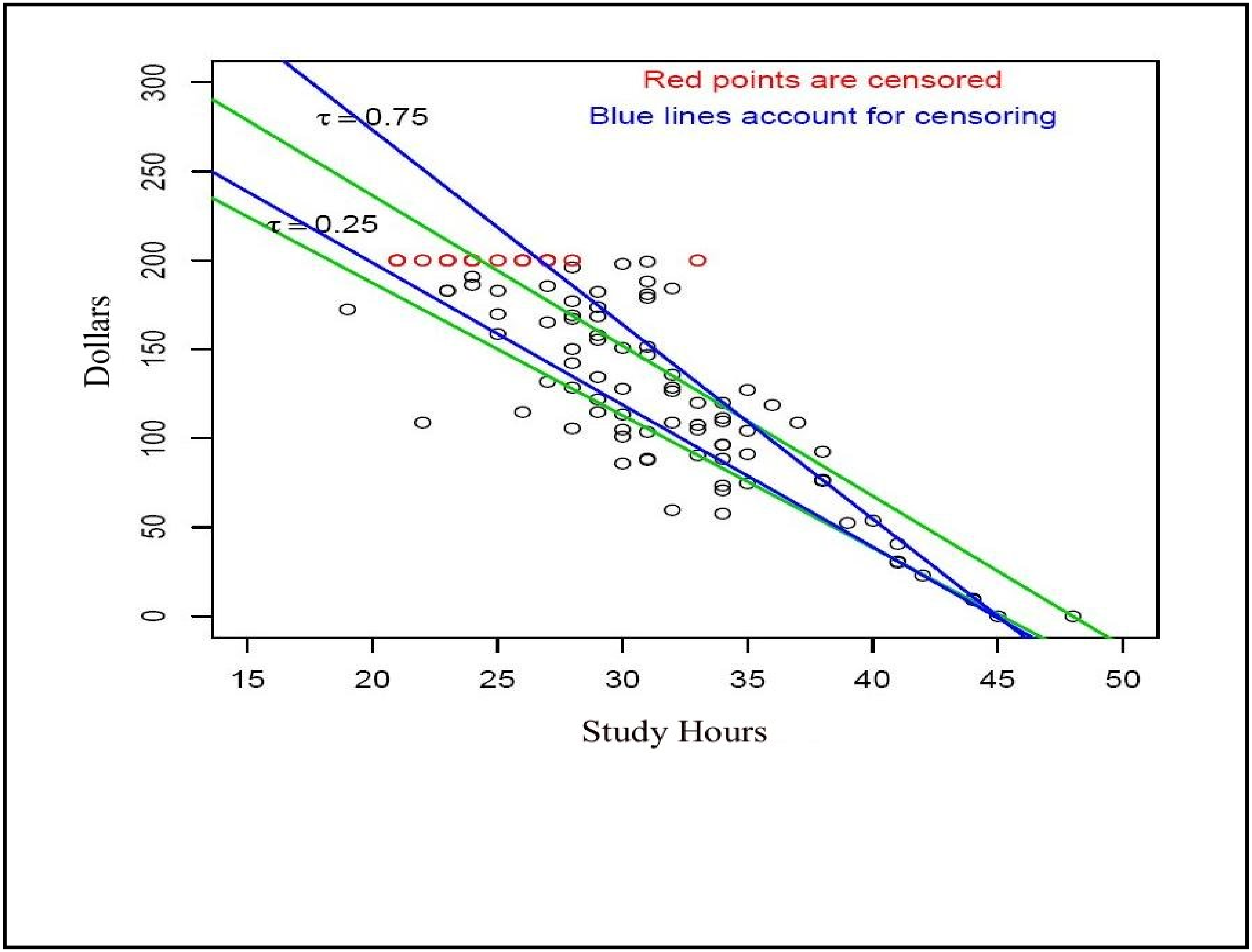
Censored quantile regression:

$$T_i = \mathbf{x}_i^T \boldsymbol{\beta}_0(\tau) + e_i(\tau), \quad i = 1, \cdots, n,$$

where $e_i(\tau)$ is the random error whose τth quantile conditional on $\mathbf{x}_i$ equals 0.

Why Censored Quantile Regression?

Example: Average Weekly Earnings v.s. Study Hours



Red points are censored
Blue lines account for censoring
τ = 0.75
τ = 0.25
Dollars
Study Hours
0
50
100
150
200
250
300
15
20
25
30
35
40
45
50

7.2 Fixed Censoring

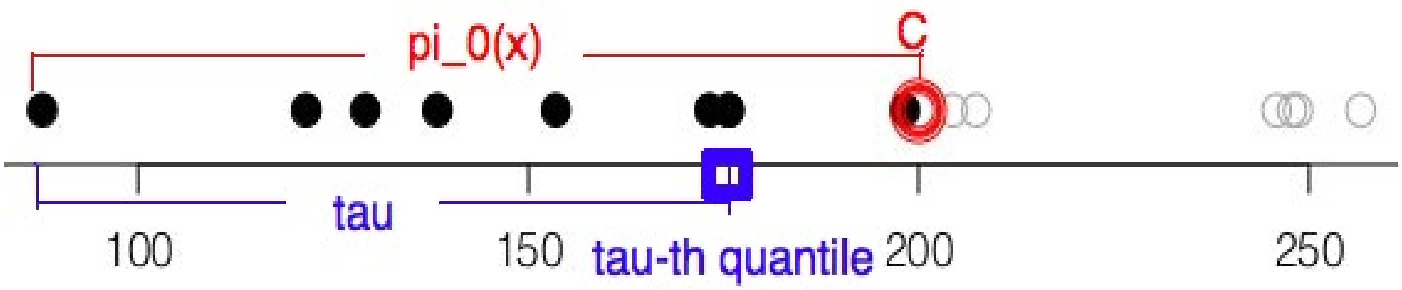
- **Fixed censoring:** the censoring times C_i are known for all observations, even for those subjects that are not censored. WLOG assume $C_i = C$.

- Examples of variables subject to fixed censoring:
  - viral load of HIV patients, antibody concentration in blood: censored due to detection limits;
  - age or salary in survey studies: censored due to top/bottom coding.

Identifiability under Censoring

- Conditional mean $E(T|X)$ is **not identifiable**.

- But the conditional quantiles $Q_\tau(T|X)$ are **identifiable** for some τ.



40% right censoring (ed) at 200.

Identifiable quantile region: $\tau \in (0, 0.6)$.

Powell's Estimator

$$Q_\tau(T|\mathbf{x}_i) = \mathbf{x_i^T}\boldsymbol{\beta_0}(\tau), \quad Y_i = \min(T_i, C)$$
$$\Rightarrow \quad Q_\tau\{Y|\mathbf{x}_i\} = \min\{\mathbf{x_i^T}\boldsymbol{\beta_0}(\tau), \mathbf{C}\}.$$

- Powell's estimator:

$$\hat{\boldsymbol{\beta}}(\tau) = argmin_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho_\tau\{Y_i - \min(C, \mathbf{x}_i^T\boldsymbol{\beta})\}.$$

- **Computational challenges**
  - **non-convex objective function;**
  - **easy to get stuck at a local minimum.**

References: Powell (1984, 1986)

7.3 Random Censoring

Assume C_i and T_i are conditionally independent given X_i.

Two iterative censored quantile regression algorithms:

- Portnoy (2003): split each censored point into two with proper weights.
- Peng and Huang (2008): use martingale-based estimating equations.

8 Applications

given by Professor Ying Wei