

# Flexible Regression

Session 2 - Introduction to Quantile Regression

Claire Miller & Tereza Neocleous

# Session outline

1. Definitions
2. Motivating examples
3. Estimation
4. Asymptotics
5. Inference
6. Nonparametric quantile regression

# What is quantile regression?

What is a quantile?

$Y$ : random variable with CDF  $F_Y(y) = P(Y \leq y)$ .

The  $\tau$ th quantile of  $Y$  is

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

$\tau$ : quantile level,  $0 < \tau < 1$ .

- ▶  $\tau = 0.25$ : first quartile
- ▶  $\tau = 0.5$ : median
- ▶  $\tau = 0.75$ : third quartile

$Q_\tau(Y)$ : **nondecreasing function** of  $\tau$ .

# Conditional quantile

## Regression setting

$Y$ : response variable

$\mathbf{x}$ :  $p$ -dimensional predictor

$F_Y(y|\mathbf{x}) = P(Y \leq y|\mathbf{x})$ : conditional CDF of  $Y$  given  $\mathbf{x}$

Then the  $\tau$ **th conditional quantile** of  $Y$  is defined as

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

# Mean vs quantile regression

- ▶ Least squares linear mean regression model:

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

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

- ▶ Linear quantile regression model:

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

$Q_\tau(Y|\mathbf{x})$  is a non-decreasing function of  $\tau$  for any given  $\mathbf{x}$ .

## Example: location-scale shift model

Consider random variables  $Y_i$ ,  $i = 1, \dots, n$  where

$$Y_i = \alpha + \mathbf{z}_i^T \boldsymbol{\beta} + (1 + \mathbf{z}_i^T \boldsymbol{\gamma}) \varepsilon_i,$$

with  $\varepsilon \stackrel{\text{i.i.d.}}{\sim} F(\cdot)$ .

Conditional quantile function:

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

- ▶  $\alpha(\tau) = \alpha + F^{-1}(\tau)$  is nondecreasing in  $\tau$ ;
- ▶  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \boldsymbol{\gamma} F^{-1}(\tau)$  may depend on  $\tau$ .

**Location shift:**  $\boldsymbol{\gamma} = \mathbf{0}$ , so that  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}$  is constant across  $\tau$ .

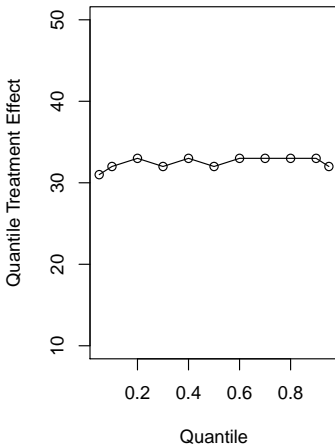
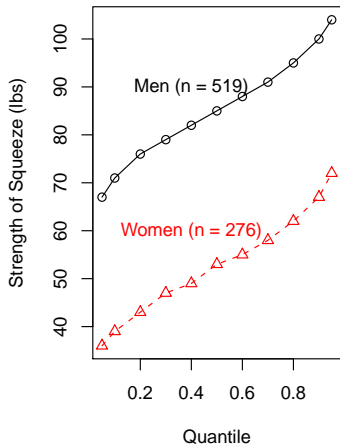
## ANTHROPOMETRIC PER-CENTILES

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

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

Subject of measurement	Age	Unit of measurement	Sex	No. of persons in the group	Values surpassed by per-cents. as below										
					95	90	80	70	60	50	40	30	20	10	5
					Values unreachd 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

# Galton's strength of squeeze data





## Quantile treatment effects

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

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

- ▶ Quantile treatment effect:

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

- ▶ Thus

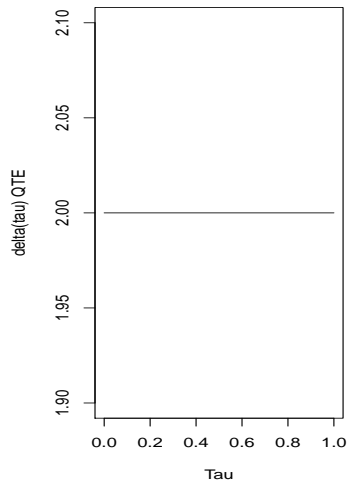
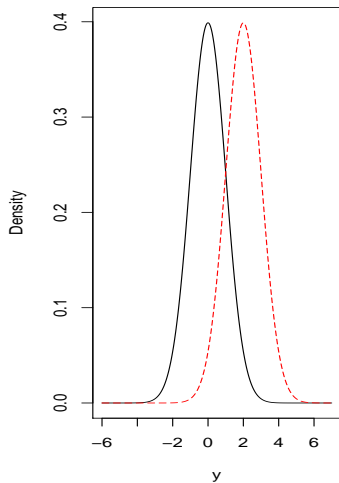
$$\Delta = \int_0^1 G^{-1}(u)du - \int_0^1 F^{-1}(u)du = \int_0^1 \delta(u)du.$$

- ▶ Equivalent quantile regression model (with binary covariate):

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

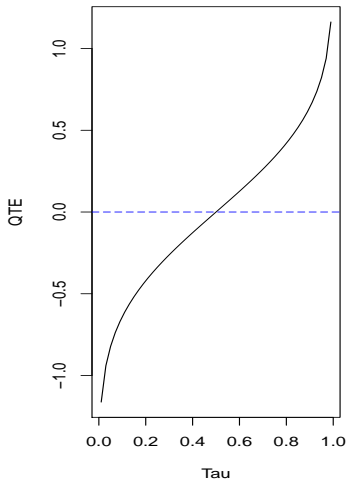
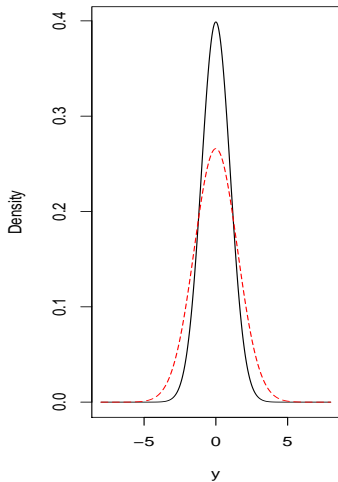
# Location shift

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

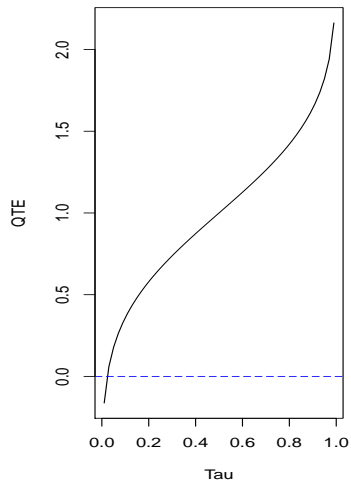
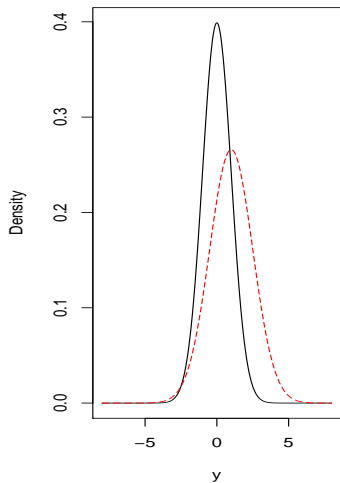


# Scale shift

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



# Location-scale shift

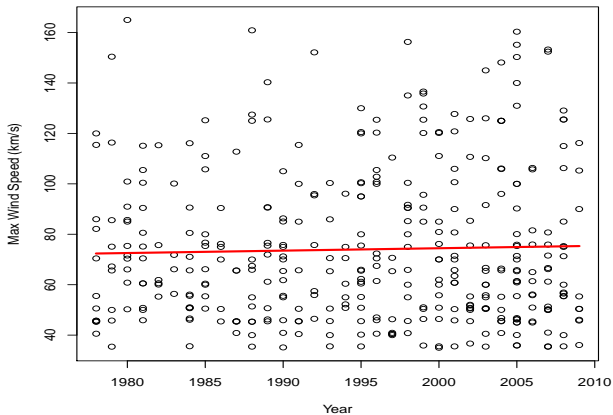


# Why quantile regression?

1. To study the impact of predictors on different quantiles of the response distribution in order to provide a complete picture of the relationship between  $Y$  and  $\mathbf{x}$ .

## Example: Tropical cyclones

- ▶  $y_i$  : max wind speeds of tropical cyclones in the North Atlantic
- ▶  $x_i$ : year 1978-2009



OLS estimate

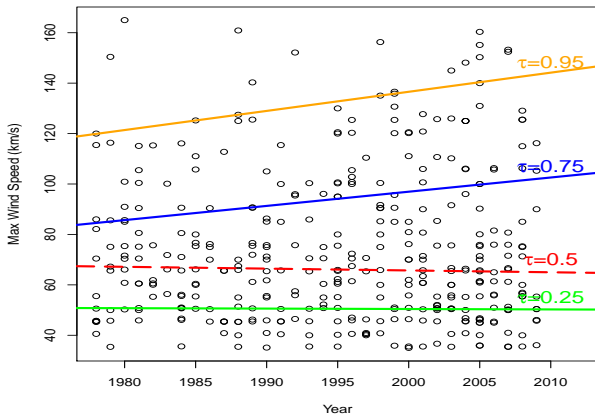
$$\hat{\beta} = 0.095$$

$p$ -value:

0.569

# Example: Tropical cyclones

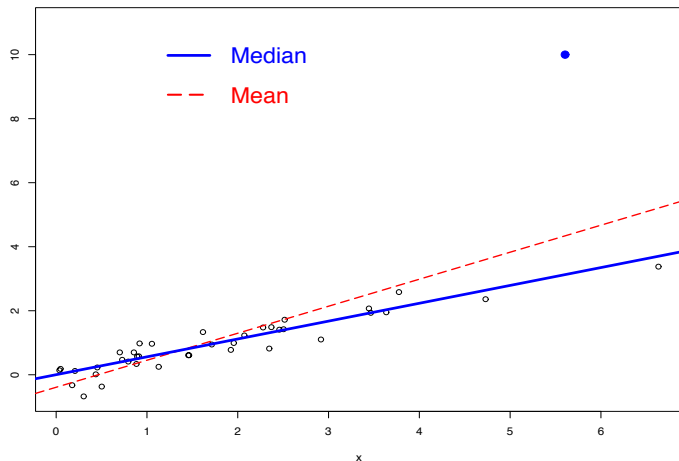
Do the **quantiles** of max speed change over time?



$\tau$	$p$ -value
0.95	0.009
0.75	0.100
0.50	0.718
0.25	0.659

# Why quantile regression?

2. It is robust to outliers in  $y$  observations.





## Why quantile regression?

- 
- 
3. It makes no distributional assumptions.

## Equivariance properties

- ▶  $\hat{\beta}(\tau; ay, \mathbf{X}) = a\hat{\beta}(\tau; y, \mathbf{X})$  for any constant  $a > 0$
- ▶  $\hat{\beta}(\tau; -ay, \mathbf{X}) = -a\hat{\beta}(1 - \tau; y, \mathbf{X})$  (**scale equivariance**)
- ▶  $\hat{\beta}(\tau; y + \mathbf{X}\gamma, \mathbf{X}) = \hat{\beta}(\tau; y, \mathbf{X}) + \gamma$  where  $\gamma \in \mathbb{R}^p$  (**regression shift**)
- ▶  $\hat{\beta}(\tau; y, \mathbf{X}A) = A^{-1}\hat{\beta}(\tau; y, \mathbf{x})$  where  $A$  is any  $p \times p$  nonsingular matrix (**reparameterisation of design**)

## Equivariance to monotone transformations

Suppose  $h(\cdot)$  is an increasing function on  $\mathbb{R}$ . Then for any variable  $Y$ ,

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

That is, the quantiles of the transformed random variable  $h(Y)$  are simply the transformed quantiles on the original scale.

This is not true in general for the mean, e.g.

$$\mathbb{E}(\log(Y)|X) \neq \log(\mathbb{E}(Y|X))$$

but

$$Q_\tau(\log(Y|X)) = \log(Q_\tau(Y|X)).$$

# Interpolation

Linear quantile regression lines exactly fit  $p$  observations (**subgradient condition**).

Which  $p$  points should be interpolated is determined by using all observations.

## Estimation of quantile regression coefficients

### Mean regression – ordinary least squares (OLS)

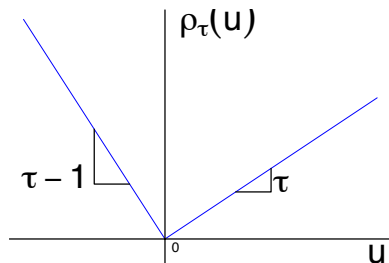
- ▶ The mean  $E(Y)$  minimises  $E\{(Y - a)^2\}$ .
- ▶ The sample mean minimises  $\sum_{i=1}^n (y_i - a)^2$ .
- ▶ The OLS estimator minimises  $\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$ .

### Median regression – least absolute deviation (LAD)

- ▶ The median  $Q_{0.5}(Y)$  minimises  $E|Y - a|$ .
- ▶ The sample median minimises  $\sum_{i=1}^n |y_i - a|$ .
- ▶ Assuming  $Q_{0.5}(y|x) = \mathbf{x}_i^T \boldsymbol{\beta}(0.5)$ ,  $\hat{\boldsymbol{\beta}}(0.5)$  can be obtained by minimising  $\sum_{i=1}^n |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$ .

## Quantile coefficient estimation

- ▶ The  $\tau$ th quantile  $Q_\tau(Y)$  minimises  $E\{\rho_\tau(Y - a)\}$ , where  $\rho_\tau(u) = u\{\tau - I(u < 0)\}$  is the quantile loss function.



- ▶ The  $\tau$ th sample quantile of  $Y$  minimises  $\sum_{i=1}^n \rho_\tau(y_i - a)$ .
- ▶ Assuming  $Q_\tau(Y|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}(\tau)$ , then  $\hat{\boldsymbol{\beta}}(\tau)$  minimises  $\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$ .

# How to minimise the objective function?

## Linear programming 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

$$y_1 \geq 0, \dots, y_m \geq 0,$$

where  $\mathbf{A}$  is an  $m \times n$  matrix,  $\mathbf{b} \in \mathbb{R}^m$ ,  $\mathbf{c} \in \mathbb{R}^n$ .

# How to minimise the objective function?

## Dual problem

$$\max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x},$$

subject to constraints

$$\mathbf{Ax} \leq \mathbf{b}$$

and

$$\mathbf{x} \geq 0.$$



## Quantile regression as a linear programming problem

$$\begin{aligned}y_i &= \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) + e_i \\ &= \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) + (u_i - v_i),\end{aligned}$$

where

$$\begin{aligned}u_i &= e_i I(e_i > 0), \\ v_i &= |e_i| I(e_i < 0).\end{aligned}$$

So

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

# Implementation in R

- ▶ Function `rq()` from `library(quantreg)` fits quantile regression models.

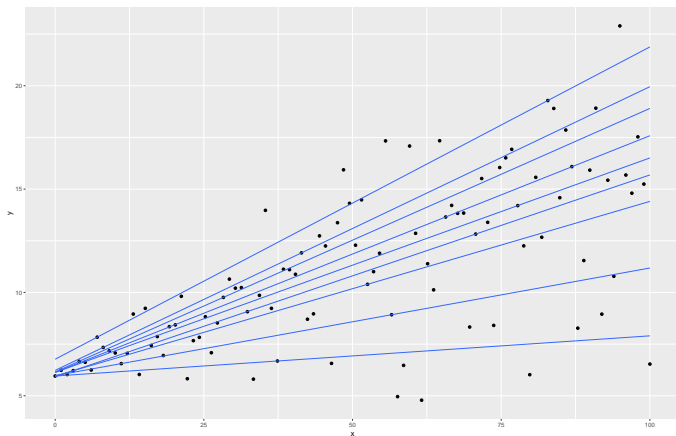
- ▶ Syntax:

```
rq(y ~ x, tau=.5, data,method=...)
```

- ▶ `method="br"` (default) implements the simplex method of Barrodale and Roberts (1974) for optimising the objective function.
- ▶ `method="fn"` implements the Frisch-Newton interior point algorithm (Portnoy and Koenker, 1997).
- ▶ `method="sfn"` implements a version of the interior point algorithm suitable for sparse design matrices (Koenker and Ng, 2003).

## Example: illustration with simulated data

```
library(quantreg)
taus <- 1:9/10
fit <- rq(y ~ x, data=dat, tau = taus)
ggplot(dat, aes(x,y)) + geom_point()
  + geom_quantile(quantiles = taus)
```



## Example: illustration with simulated data

```
> fit <- rq(y~x, data=dat, tau=.5)
> summary(fit)
```

```
Call: rq(formula = y ~ x, tau = 0.5, data = dat)
```

```
tau: [1] 0.5
```

```
Coefficients:
```

	coefficients	lower bd	upper bd
(Intercept)	6.13147	5.91573	6.42189
x	0.10376	0.09776	0.11575

# Statistical properties

Coefficient estimator

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

## Consistency

Under regularity conditions A1 and A2(i) (see next slide)

$$\hat{\beta}(\tau) \xrightarrow{P} \beta(\tau).$$

# Statistical properties

## 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  $\infty$  at  $\xi_i(\tau) = Q_\tau(Y|\mathbf{x}_i)$ .
- A2. There exist positive definite matrices  $D_0$  and  $D_1$  such that
- (i)  $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T = D_0$ ;
  - (ii)  $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n f_i(\xi_i(\tau)) \mathbf{x}_i \mathbf{x}_i^T = D_1(\tau)$ ;
  - (iii)  $\max_{i=1, \dots, n} \|\mathbf{x}_i\| = o(n^{\frac{1}{2}})$ .

# Statistical properties

## Asymptotic normality

Under Conditions A1 and A2

$$\sqrt{n} \left( \hat{\beta}(\tau) - \beta(\tau) \right) \xrightarrow{d} N \left( 0, \tau(1 - \tau) D_1^{-1} D_0 D_1^{-1} \right).$$

## Simplification in the case of i.i.d. errors

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

where  $f_i(\xi_i(\tau)) = f_\varepsilon(0)$ .

# Inference

- ▶ **Idea:** use asymptotic normality results to perform Wald-type hypothesis tests and construct confidence intervals.
- ▶ **Problem:** Asymptotic covariance matrix involves the unknown densities  $f_i(\mathbf{x}_i^T \boldsymbol{\beta}(\tau))$  in non-*i.i.d.* settings, and  $f_\varepsilon(0)$  in *i.i.d.* settings.

**How do we estimate these?**



## Estimation in i.i.d. setting

### Sparsity parameter

$s(\tau) = \frac{1}{f(F^{-1}(\tau))}$  (derivative of the quantile function  $F^{-1}(t)$  with respect to  $t$ )

### Difference quotient estimator (Siddiqui,1960)

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

where

- ▶  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ ,
- ▶  $\hat{F}_n^{-1}(t|\bar{\mathbf{x}})$  is the estimated  $t$ th conditional quantile of  $Y$  given  $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$ .

## Estimation in non-i.i.d. settings

### Estimation of $D_1(\tau)$

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

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

where  $\xi_i(\tau) = Q_\tau(Y|\mathbf{x}_i)$ .

**“Hendricks-Koenker sandwich”**

# Implementation in R

```
> # Assuming iid errors:  
> summary.rq(fit, se="iid")  
  
> # Hendricks-Koenker sandwich:  
> summary.rq(fit, se="nid") # assuming non-iid errors  
tau: [1] 0.5
```

Coefficients:

	Value	Std. Error	t value	Pr(> t )
(Intercept)	6.13147	0.17754	34.53611	0.00000
x	0.10376	0.00888	11.67973	0.00000

```
> # Based on Powell kernel estimator  
> summary.rq(fit, se="ker")
```

# Rank score test

- ▶ Model:  $Q_\tau(Y|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) + \mathbf{z}_i^\top \boldsymbol{\gamma}(\tau)$
- ▶ Hypotheses:  $H_0 : \boldsymbol{\gamma}(\tau) = 0$  vs  $H_1 : \boldsymbol{\gamma}(\tau) \neq 0$   
where  $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$  and  $\boldsymbol{\gamma}(\tau) \in \mathbb{R}^q$ .
- ▶ Score function:

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

where

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

## Rank score test

- ▶ Under  $H_0$ , as  $n \rightarrow \infty$ ,

$$S_n = AN(0, M_n^{\frac{1}{2}}),$$

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

- ▶ Then the rank-score test statistic

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

- ▶ In *i.i.d.* settings  $\mathbf{z}^* = (\mathbf{z}_i^*) = \{\mathbf{I} - \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top\} \mathbf{z}$  and  $M_n = \tau(1 - \tau)n^{-1} \sum_{i=1}^n \mathbf{z}_i^* \mathbf{z}_i^{*\top}$  – no need to estimate the nuisance parameters  $f_i\{Q_\tau(Y|\mathbf{x}_i, z_i)\}$ .
- ▶ The rank score test can be inverted to give confidence intervals.

## Implementation in R

The rank score method is the default method for standard error and confidence interval estimation in `library(quantreg)`:

```
> # assuming iid errors
> summary.rq(fit, se="rank", alpha=0.05, iid=TRUE)
> # assuming non-iid errors
> summary.rq(fit, se="rank", alpha=0.05, iid=FALSE)
```

```
tau: [1] 0.5
```

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	6.13147	5.81521	6.54475
x	0.10376	0.08918	0.11880

# Bootstrap methods

- ▶ An alternative approach is to use bootstrap for standard error estimation
- ▶ Options include:
  - ▶ **residual bootstrap**
  - ▶ **paired bootstrap**
  - ▶ **Markov chain marginal bootstrap (MCMB)**
  - ▶ ...
- ▶ See `boot.rq()` in `library(quantreg)`

```
> summary.rq(fit, se="boot", alpha=0.05) # default: paired  
tau: [1] 0.5
```

Coefficients:

	Value	Std. Error	t value	Pr(> t )
(Intercept)	6.13147	0.20251	30.27766	0.00000
x	0.10376	0.00772	13.43691	0.00000

# Nonparametric quantile regression

- ▶ The ideas of
  - ▶ **local polynomial models**,
  - ▶ **regression splines**,
  - ▶ **penalised splines**,

introduced earlier, can be applied to quantile regression.

- ▶ Decisions about the order of the spline, number of knots or penalty parameter need to be made.

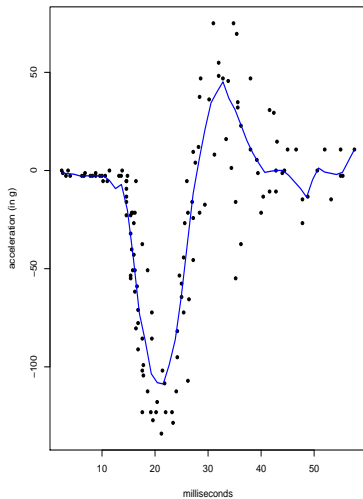
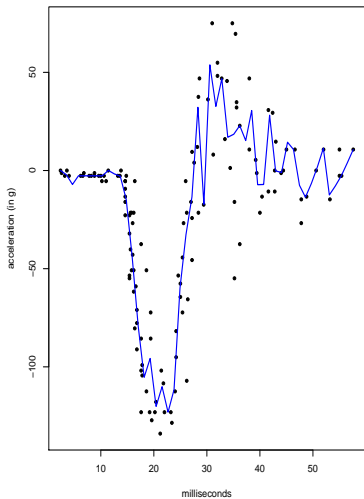


## Example: motorcycle data

- ▶ Locally linear approach using the `lprq` function from `library(quantreg)`.
- ▶ This function computes a quantile regression fit at each of  $m$  equally spaced  $x$ -values over the support of the observed  $x$  points.
- ▶ The value of the smoothing parameter (bandwidth  $h$ ) must be provided.
- ▶ In R:

```
> library(MASS) # to get the mcycle data
> fit1 <- lprq(mcycle$times,mcycle$accel,h=.5,tau=0.5)
> fit2 <- lprq(mcycle$times,mcycle$accel,h=2,tau=0.5)
```

# Local linear median regression fit for the motorcycle data with $h=0.5$ and $h=2$

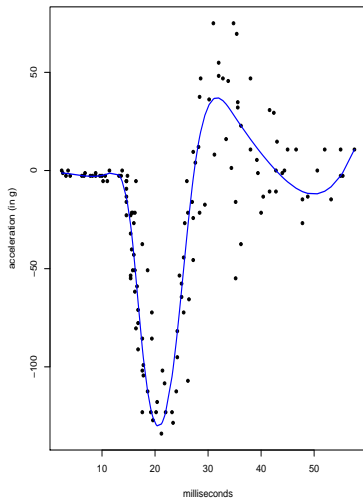
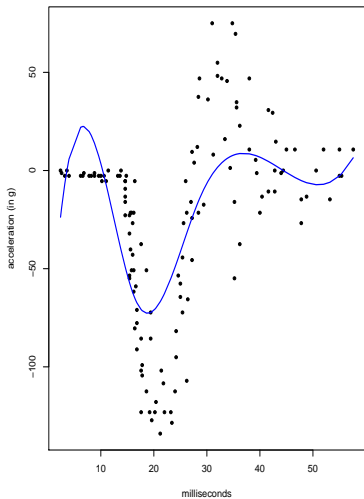


## Example: motorcycle data

- ▶ B-splines can be implemented using the function `bs()` in the package `splines` in R.
- ▶ Here we control the level of smoothing via the degrees of freedom.  

```
> fit3 <- rq(accel~bs(times,df=5),tau=0.5, data=mcycle)
> fit4 <- rq(accel~bs(times,df=10),tau=0.5, data=mcycle)
```

# Median regression fit using cubic B-splines with $df=5$ and $df=10$ for the motorcycle data

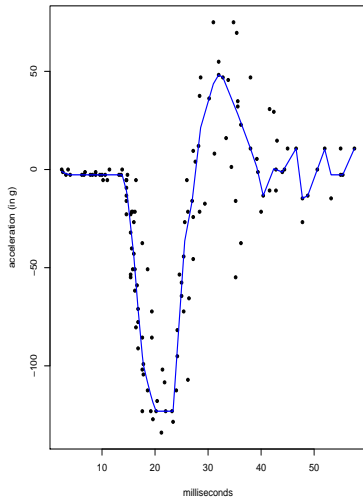
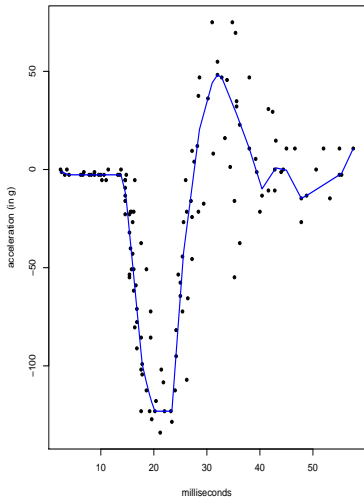


## Example: motorcycle data

- ▶ Quantile smoothing splines using a roughness penalty can be implemented via the `rqss()` function in `library(quantreg)` in R.
- ▶ This function is quite flexible and allows specification of monotonicity and convexity constraints.
- ▶ Penalty parameter  $\lambda$  has to be specified by the user (default value is `lambda=1`).
- ▶ In R:

```
> fit5 <- rqss(accel~qss(times,constraint="N", lambda=1),
               tau=0.5, data=mcycle)
> fit6 <- rqss(accel~qss(times,constraint="N", lambda=0.5),
               tau=0.5, data=mcycle)
```

Median regression fit for the motorcycle data using quantile smoothing splines with penalty  $\lambda = 1$  and  $\lambda = 0.5$ .



## Remarks

- ▶ Spline methods are better than local linear methods in general.
- ▶ All methods require decisions to be made about the degree of smoothing to be applied.
- ▶ Quantile crossing is an issue in general, and even more so with nonparametric quantile regression, especially for  $\tau$  near 0 or 1.

# Summary

## Quantile regression

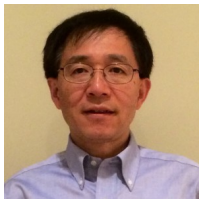
- ▶ Quantiles and quantile regression
- ▶ Why use quantile regression?
- ▶ How to fit quantile regression models in R
- ▶ Splines for nonparametric quantile regression



# Aknowledgements



Prof. Judy H. Wang,  
GWU



Prof. Xuming He,  
U Michigan



Prof. Roger Koenker,  
U Illinois