apts.ac.uk
Academy for PhD Training in Statistics

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

The role of linguistic diversity in the prediction of early reading comprehension: A **quantile regression** approach

LJ van den Bosch, E Segers… - Scientific Studies of …, 2019 - Taylor & Francis

Using classical and **quantile regression** analyses, we investigated whether predictor variables for early reading comprehension differed depending on language background and ability level in a mixed group of 161 monolingual (L1) and bilingual (L2) children in second …

☆ 99 Cited by 1 Related articles All 5 versions

Variation across price segments and locations: A comprehensive **quantile regression** analysis of the Sydney housing market

SR Waltl - Real Estate Economics, 2019 - Wiley Online Library

Standard house price indices measure average movements of average houses in average locations belonging to an average price segment and hence obscure spatial and cross-sectional variation of price appreciation rates even within a single metropolitan area. This …

☆ 99 Cited by 13 Related articles All 4 versions »

[HTML] **Quantile regression** analysis reveals widespread evidence for gene-environment or gene-gene interactions in myopia development

A Pozarickij, C Williams, PG Hysi… - Communications …, 2019 - nature.com

A genetic contribution to refractive error has been confirmed by the discovery of more than 150 associated variants in genome-wide association studies (GWAS). Environmental factors such as education and time outdoors also demonstrate strong associations. Currently …

☆ 99 Related articles All 9 versions

[HTML] Asymmetric effects of monetary policy on firm scale in China: A **quantile regression** approach

L Fang, L He, Z Huang - Emerging Markets Review, 2019 - Elsevier

This study explores asymmetric effects of monetary policy on firm scale at different firm size levels. We find that Chinese firms respond to raising benchmark lending interest rates and deposit reserve requirements by decreasing their scales. Our **quantile regression** results …

☆ 99 Related articles All 4 versions

[HTML] Foreign exchange interventions in Brazil and their impact on volatility: A **quantile regression** approach

AP Viola, MC Klotzle, ACF Pinto… - … in International Business …, 2019 - Elsevier

This work aims to analyze the interventions conducted by the Central Bank of Brazil in the Brazilian foreign exchange market from 2003 to 2014. For this purpose, we use **quantile regression** analysis and some of its new formulas to examine the effects of government …

☆ 99 Cited by 1 Related articles All 4 versions

Differential effects of unemployment on depression in people living with HIV/AIDS: a **quantile regression** approach

C Zeng, Y Guo, YA Hong, S Gentz, J Zhang, H Zhang… - AIDS care, 2019 - Taylor & Francis

Unemployment is associated with depression in people living with HIV (PLWH). However, few studies have examined the effects of unemployment on PLWH with different levels of depression. The current study explores the plausible differential effects of unemployment on …

☆ 99 Related articles All 4 versions

What is quantile regression?

What is a quantile?
Y: random variable with CDF $F_Y(y) = P(Y \leq y)$.
The **τth quantile** of Y is

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

τ: 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 τ.

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 **τth 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:

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

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

- Linear quantile regression model:

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

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

Example: location-scale shift model

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

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

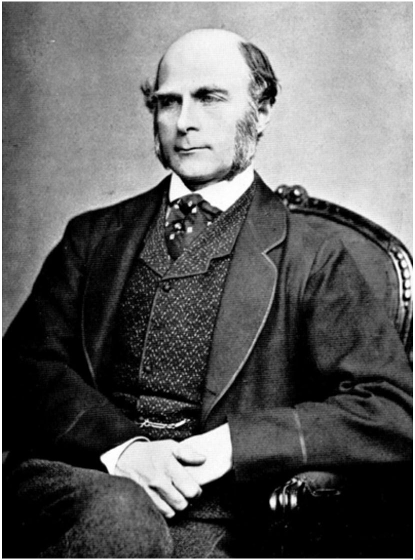
with $\varepsilon \overset{\text{i.i.d}}{\sim} F(\cdot)$.
Conditional quantile function:

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

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

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

Galton's strength of squeeze data



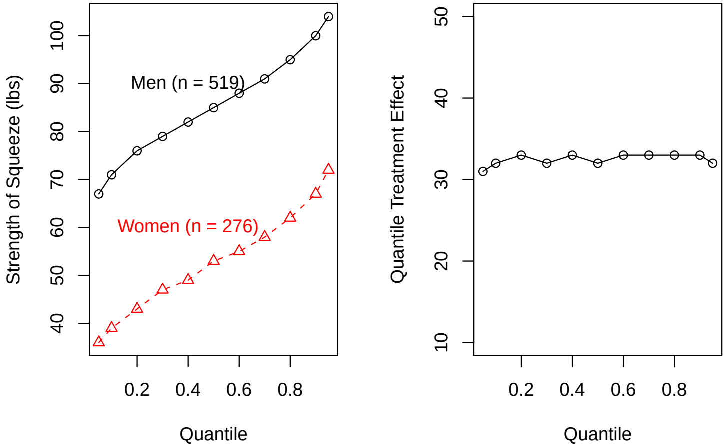
ANTHROPOMETRIC PER-CENTILES

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

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

Subject of measurement	Age	Unit of measurement	Sex	No. of persons in the group	Values surpassed by per-cents as below: 95	90	80	70	60	50	40	30	20	10	5
					Values unreached by per-cents. as below: 5	10	20	30	40	50	60	70	80	90	95
Height, standing, without shoes ...	23–51	Inches	M.	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:

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

- Quantile treatment effect:

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

- Thus

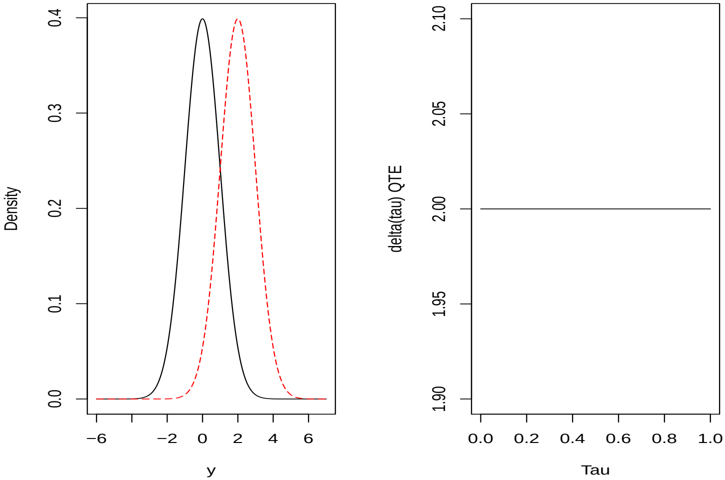
$$\Delta = \int_0^1 G^{-1}(u) du - \int_0^1 F^{-1}(u) du = \int_0^1 \delta(u) du.$$

- Equivalent quantile regression model (with binary covariate):

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

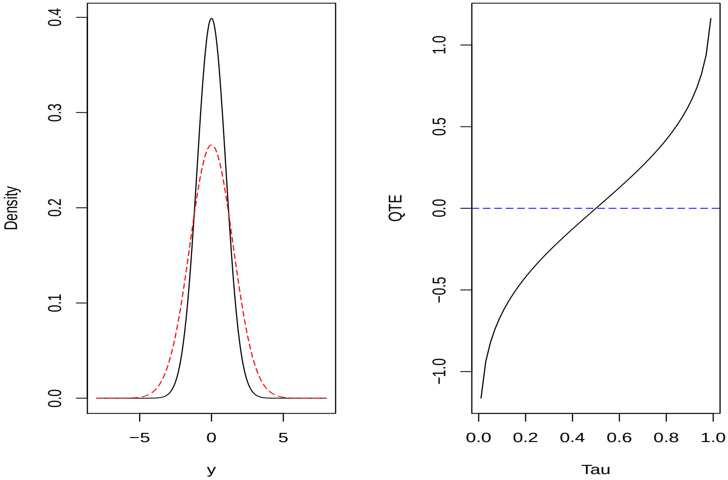
Location shift

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

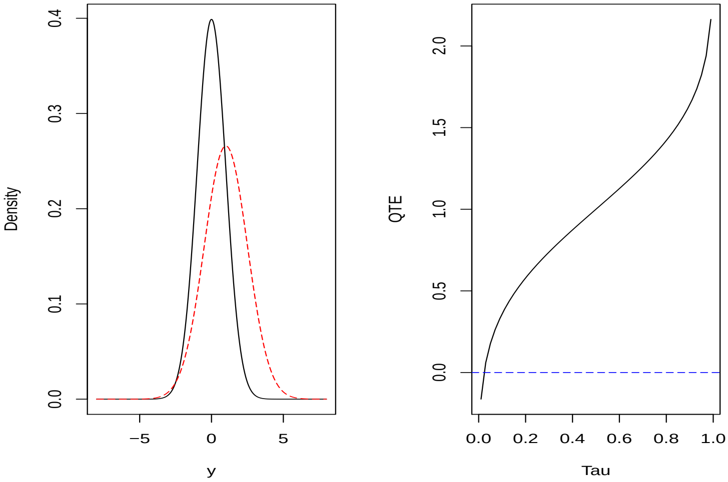


Scale shift

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



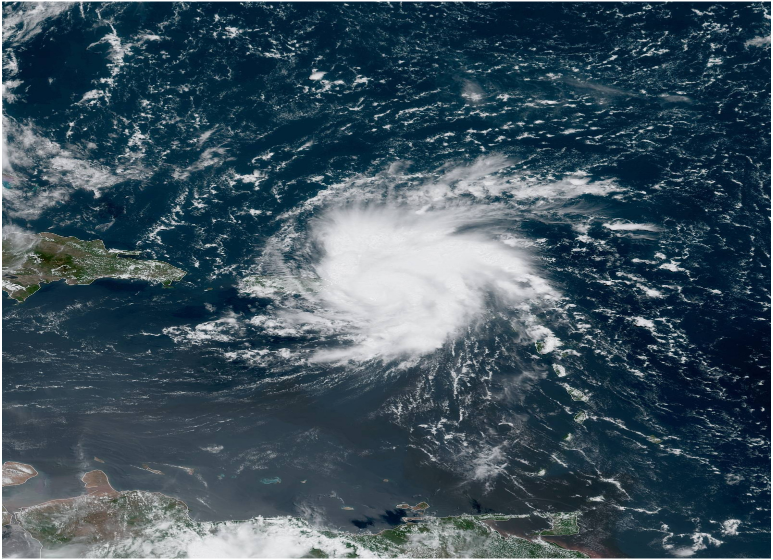
Location-scale shift



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 $\mathbf{x}$.

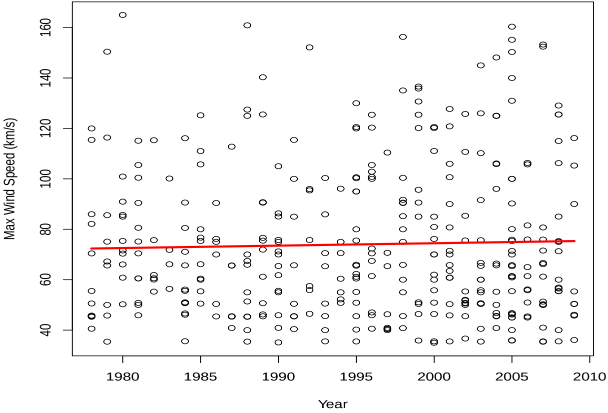
Example: Tropical cyclones



Hurricane Dorian 2019

Example: Tropical cyclones

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

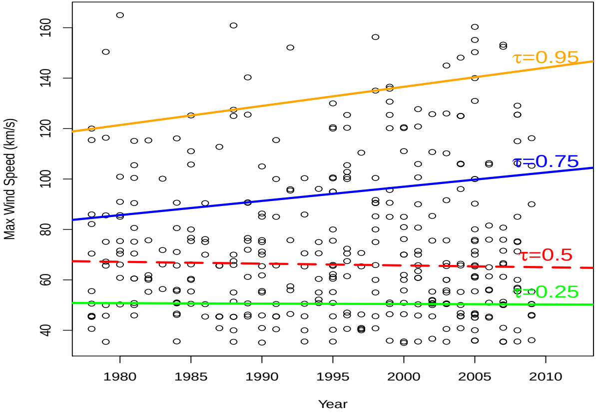


OLS estimate $\hat{\beta}$=0.095

p-value: 0.569

Example: Tropical cyclones

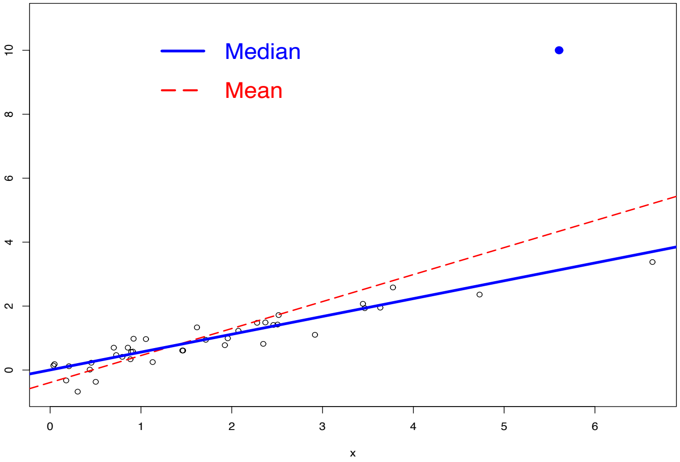
Do the **quantiles** of max speed change over time?



τ	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. (*E.g.* income distribution.)



Why quantile regression?

3. It makes no distributional assumptions.

Equivariance properties

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

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

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)

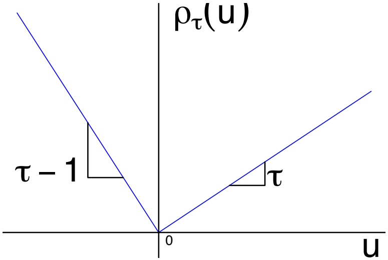
- The mean $E(Y)$ minimises $E\{(Y-a)^2\}$.
- The sample mean minimises $\sum_{i=1}^{n}(y_i-a)^2$.
- The OLS estimator minimises $\sum_{i=1}^{n}(y_i-\mathbf{x}_i^{\mathsf{T}}\beta)^2$.

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 $Q_{0.5}(y|x)=\mathbf{x}_i^{\mathsf{T}}\beta(0.5)$, $\hat{\beta}(0.5)$ can be obtained by minimising $\sum_{i=1}^{n}|y_i-\mathbf{x}_i^{\mathsf{T}}\beta|$.

Quantile coefficient estimation

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

How to minimise the objective function?

Linear programming problem

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

subject to the constraints

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

and

$$y_1 \geq 0, \cdots, y_m \geq 0,$$

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

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

subject to constraints

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}$$

and

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

Quantile regression as a linear programming problem

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

where

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

So

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

Implementation in R

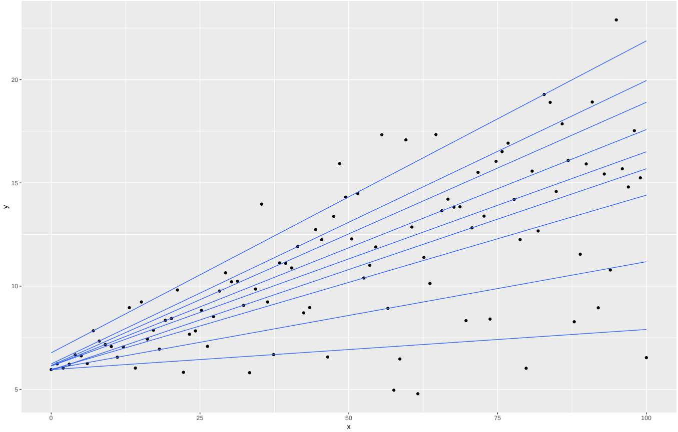
- Function `rq()` from `library(quantreg)` fits quantile regression models.
- Syntax:

```
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

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

Consistency

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

$$\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{p} \boldsymbol{\beta}(\tau).$$

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

A2. There exist positive definite matrices D_0 and D_1 such that

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

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

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

Statistical properties

Asymptotic normality

Under Conditions A1 and A2

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

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

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

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 $f_i(\mathbf{x}_i^\top \beta(\tau))$ 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

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

- $h_n \to 0$ as $n \to \infty$,
- $\hat{F}_n^{-1}(t|\bar{\mathbf{x}})$ is the estimated tth 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 τ.
- 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

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

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: $Q_\tau(Y|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}(\tau) + \mathbf{z}_i^\mathsf{T}\boldsymbol{\gamma}(\tau)$
- Hypotheses: $H_0 : \boldsymbol{\gamma}(\tau) = 0 \quad \text{vs} \quad H_1 : \boldsymbol{\gamma}(\tau) \neq 0$

  where $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$ and $\boldsymbol{\gamma}(\tau) \in \mathbb{R}^q$.
- Score function:

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

  where

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

Rank score test

- Under H_0, as $n \to \infty$,

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

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

- Then the rank-score test statistic

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

- In *i.i.d.* settings $\mathbf{z}^* = (\mathbf{z}_i^*) = \{\boldsymbol{I} - \mathbf{x}(\mathbf{x}^{\mathsf{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathsf{T}}\}\mathbf{z}$ and $M_n = \tau(1-\tau) n^{-1} \sum_{i=1}^{n} \mathbf{z}_i^* \mathbf{z}_i^{*\mathsf{T}}$ – 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**,

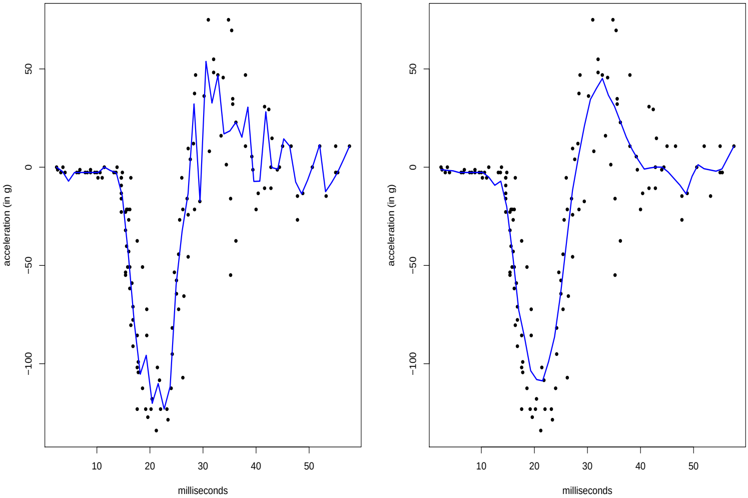
  introduced earlier, can be applied to quantile regression.
- Decisions about the order of the spline, number of knots or penalty parameter need to be made.

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`

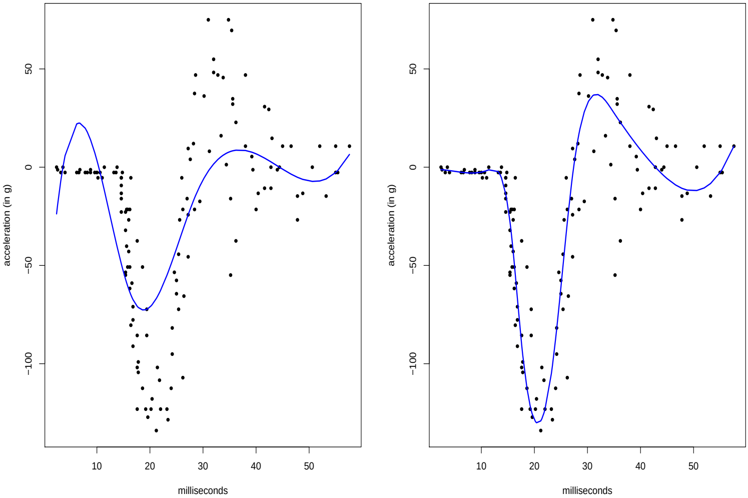


Example: motorcycle data

- 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

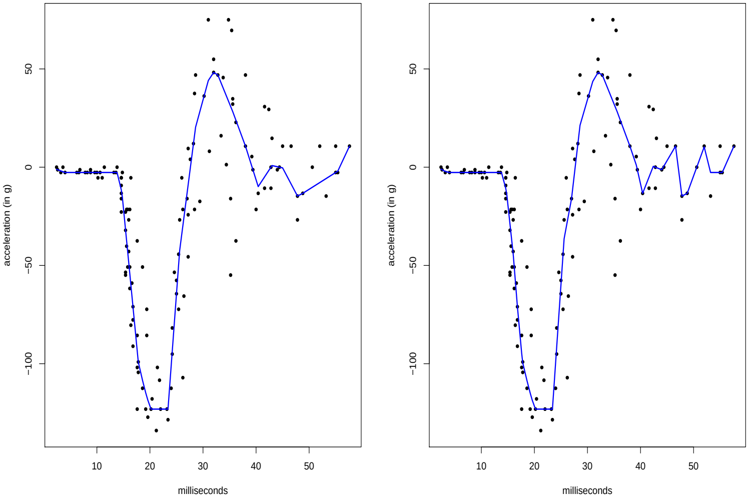


Example: motorcycle data

- 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 λ 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 τ near 0 or 1.

Example: BMI distribution

Modelling Obesity in Scotland

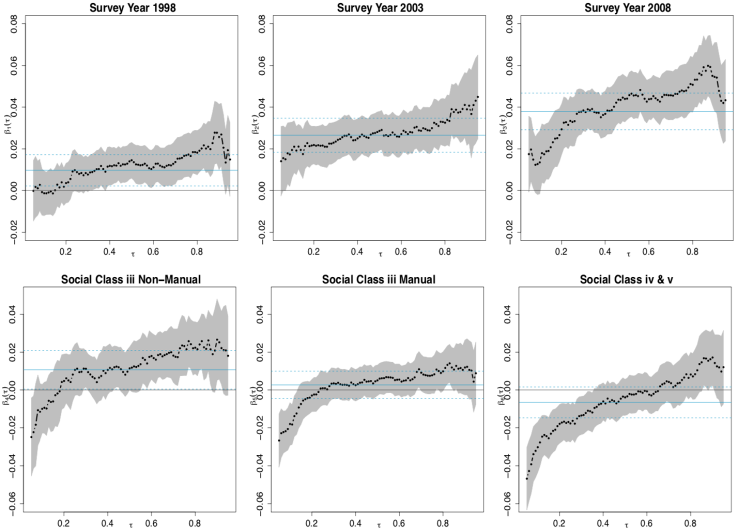
Gary Napier[1] and Tereza Neocleous

Scottish Health Survey: 1995, 1998, 2003 and 2008

$$
\begin{aligned}
Q_{\log(\mathrm{BMI})}(\tau|\mathrm{X}) = \alpha_0(\tau) &+ \sum_i \beta_i(\tau)(\mathrm{year})_i \\
&+ \sum_j \gamma_j(\tau)(\mathrm{social\ class})_j \\
&+ g_\tau(\mathrm{age})
\end{aligned}
$$

$g_\tau(\cdot)$ is a nonlinear function of age, approximated by a linear combination of cubic B-spline basis functions with fixed knots at age 35 and 49 (the 33rd and 66th percentiles of the age distribution)

Example: BMI distribution



Example: BMI distribution – year effect

log(BMI) as a function of year:

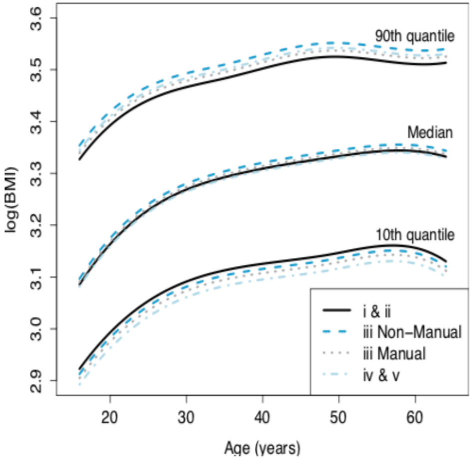
- No change in log(BMI) is observed between 1995 and 1998 at the lower quantiles, but as τ approaches 0.5 (median) an increase is revealed, which is at its largest at the upper quantiles.
- An increase in log(BMI) is observed between 1995 and 2003/2008 across the entire distribution, with log(BMI) increasing with increasing values of τ. The increase in log(BMI) is greater with each subsequent survey year, which can be seen from the upward shift on the y-axis.

Example: BMI distribution – social class effect

log(BMI) as a function of social class:

- At the bottom of the distribution, log(BMI) is lower for each social class than for social classes i & ii (baseline).
- As τ approaches 0.5 no discernible difference in log(BMI) is found between each social class and social classes i & ii.
- At the upper quantiles log(BMI) is generally higher than baseline, but not always significantly so.
- Changes in sign of the regression coefficient across the distribution highlight the benefits of quantile regression, as such fluctuations cannot be detected by least squares regression.

Example: BMI distribution – age effect



Example: BMI distribution – age effect

log(BMI) as a function of age:

- The rate of increase in log(BMI) with age is at its greatest in the early years of adulthood and gradually diminishes before starting to decrease at around 60 years of age.
- This increase is most prominent at the upper quantiles, where the separation between social classes is also at its greatest.
- As the data is not longitudinal, we cannot distinguish between generational effects and ageing.

Summary

Quantile regression

- Quantiles and quantile regression
- Why use quantile regression? Reasons and examples
- How to fit quantile regression models in R
- How to fit nonparametric quantile regression models using splines
- More examples in the lab

Aknowledgements

Prof. Judy H. Wang, GWU

Prof. Xuming He, U Michigan

Prof. Roger Koenker, U Illinois