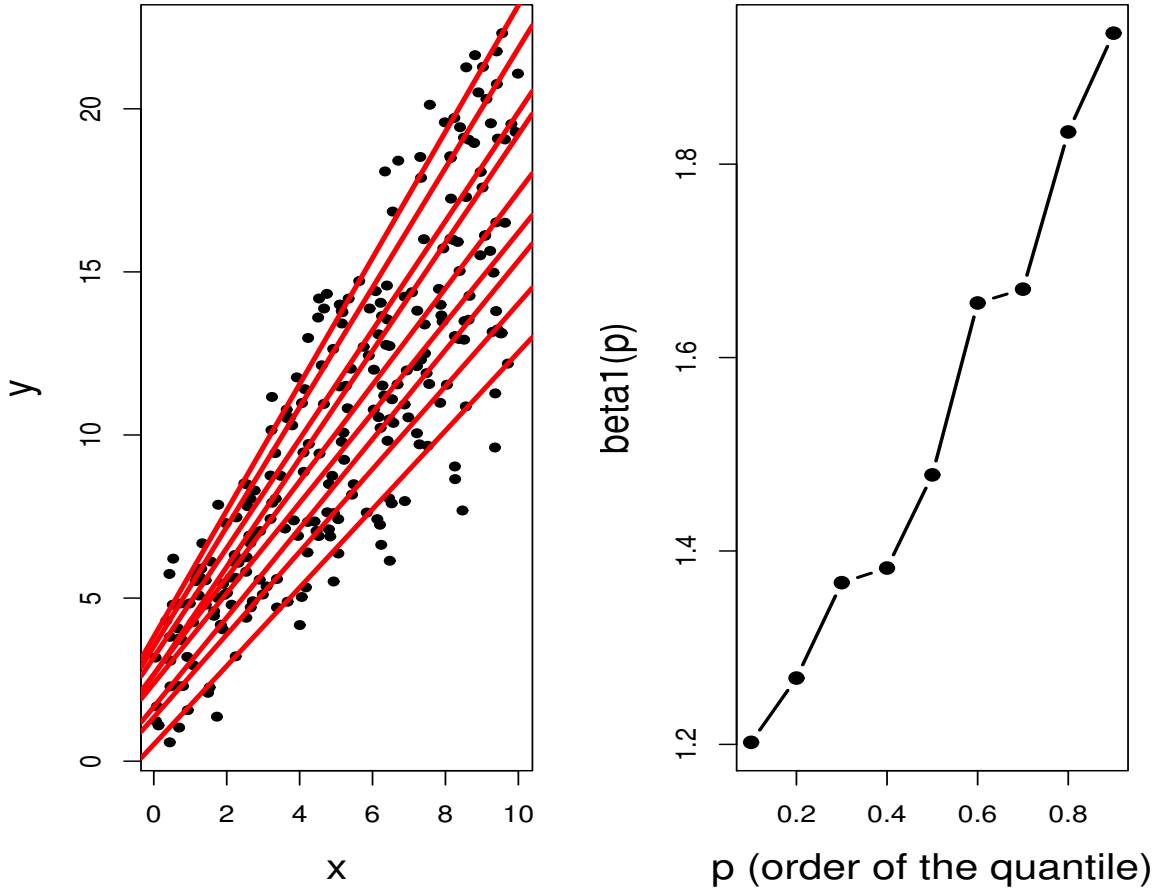
Quantile regression coefficients modeling

Paolo Frumento Matteo Bottai

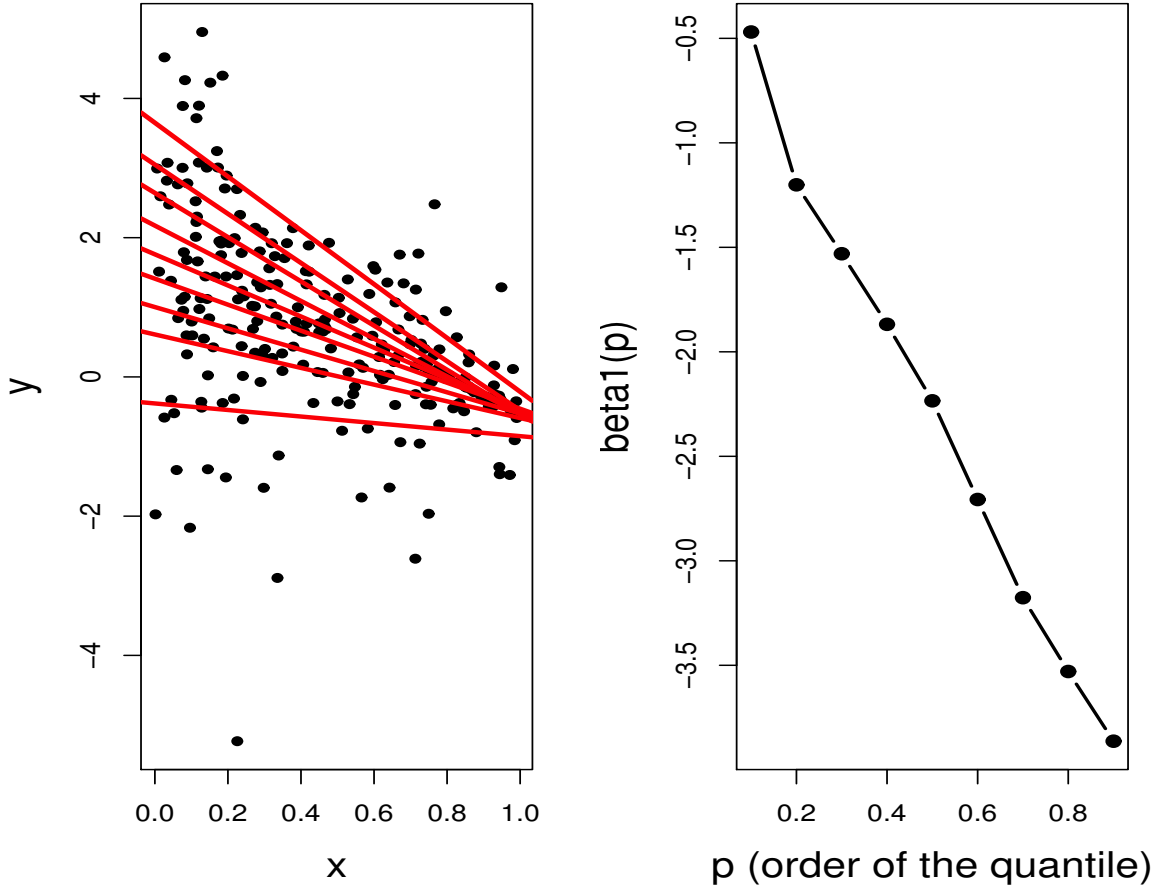
Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet
Stockholm, Sweden

Quantile regression

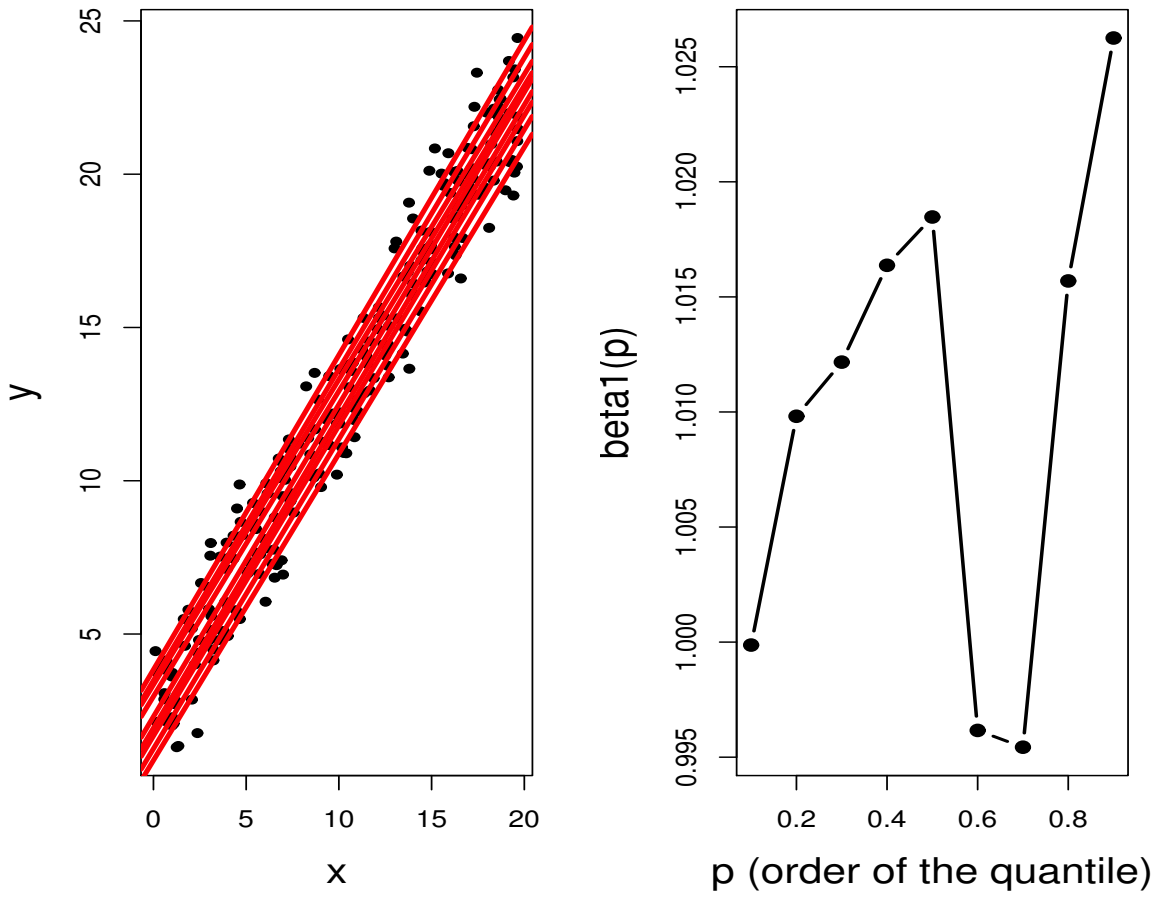
$$Q(p \mid x) = \beta_0(p) + \beta_1(p)x$$



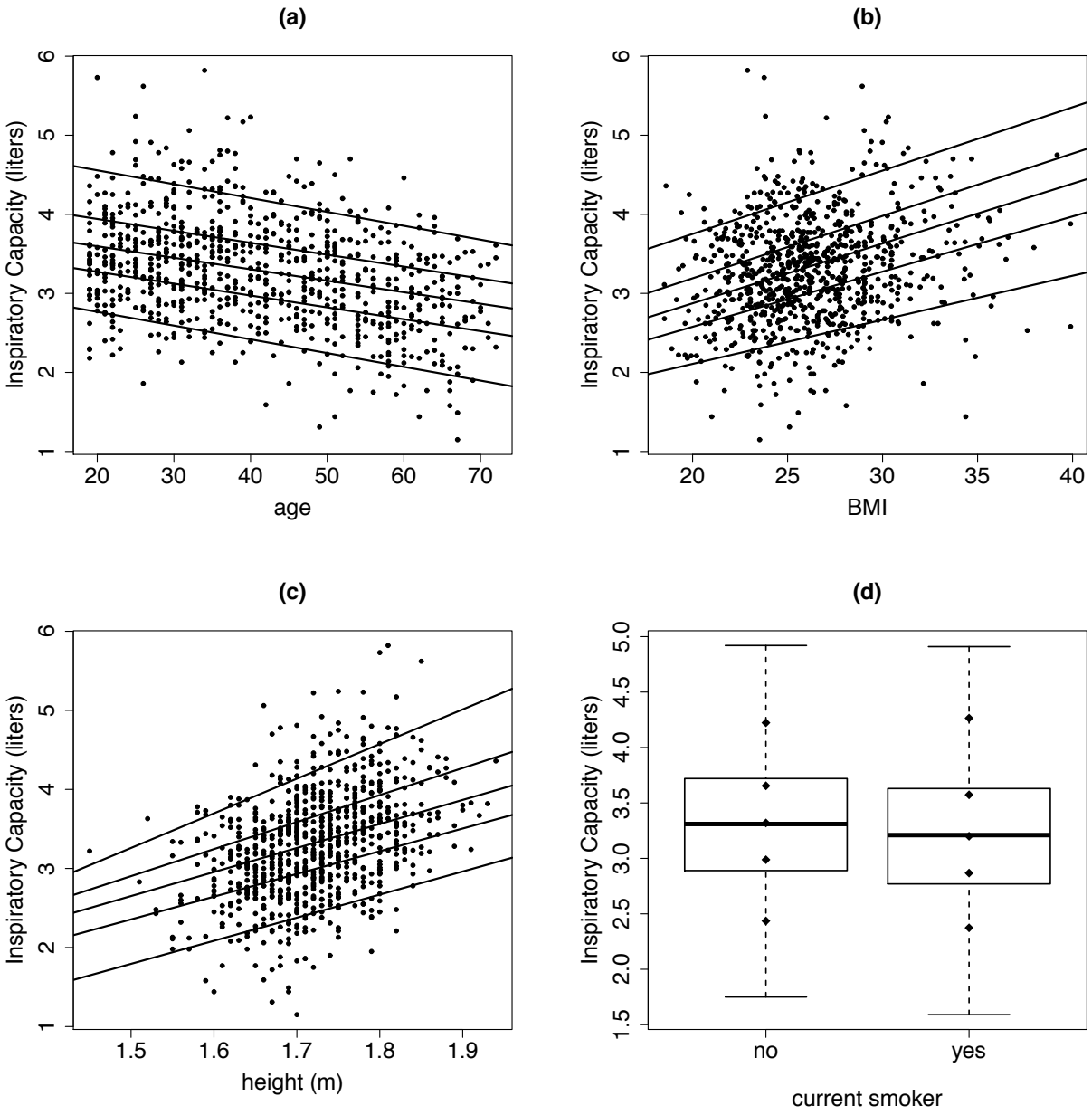
Quantile regression (2)



Quantile regression (3)



The data


(a)
Inspiratory Capacity (liters)
age
(b)
Inspiratory Capacity (liters)
BMI
(c)
Inspiratory Capacity (liters)
height (m)
(d)
Inspiratory Capacity (liters)
no
yes
current smoker

The problem

We estimated ninety-nine quantiles $(0.01, 0.02, \ldots, 0.99)$ of inspiratory capacity, a measure of lungs' volume, conditional on age, height, body mass index, and indicator of smokers.

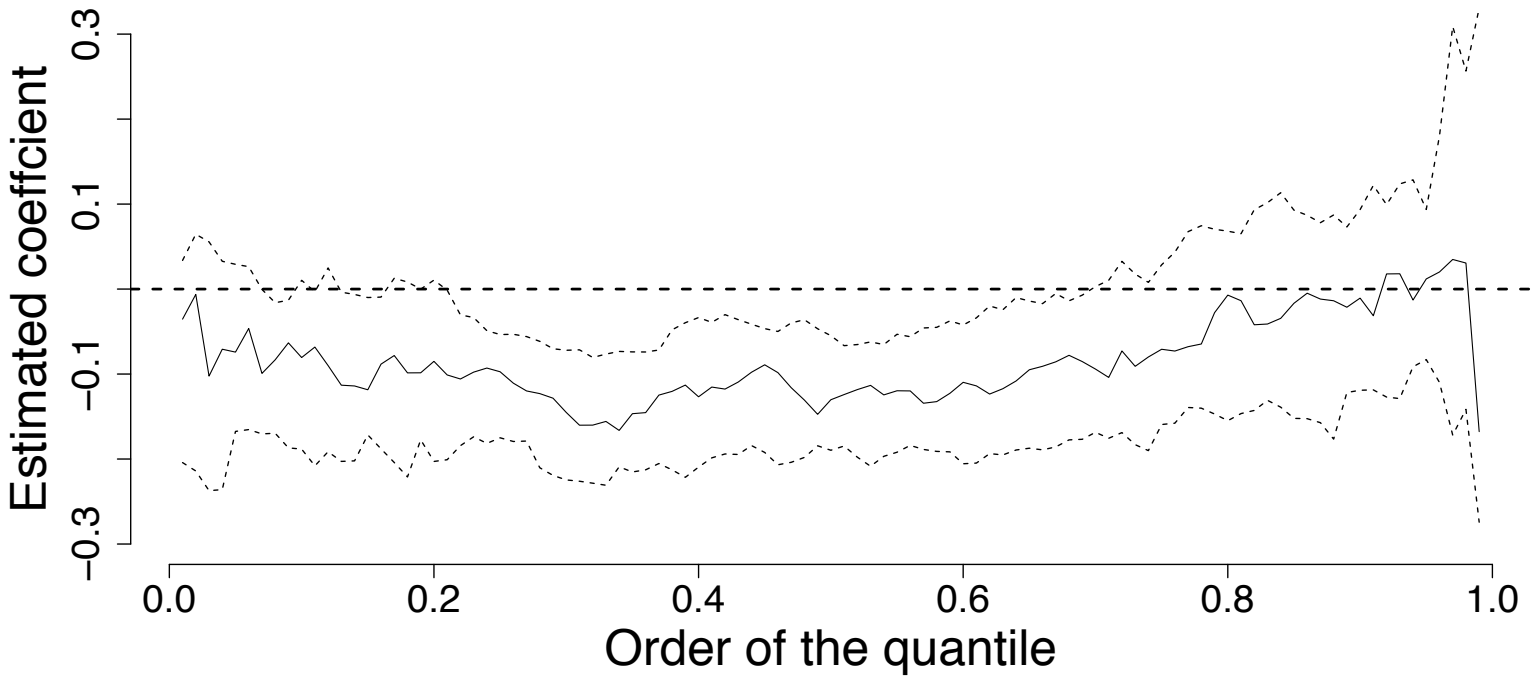


Figure: Regression coefficient of the smoking indicator at various quantiles, with pointwise 95% confidence intervals.

- Difficult to interpret
- Inefficient
- Instinctive visual interpolation

Possible solutions:

- Smoothing
- Modeling!

Quantile regression coefficients modeling (QRCM)

Linear effect of covariates on the quantile function:

$$Q(p \mid \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}(p)$$

A parametric model for $\boldsymbol{\beta}(p) = \{\beta_1(p), \ldots, \beta_q(p)\}$:

$$\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j1} b_1(p) + \ldots + \theta_{jk} b_k(p)$$

In matrix form:

$$\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta} \boldsymbol{b}(p)$$

where

$$\boldsymbol{b}(p) = [b_1(p), \ldots, b_k(p)]^{\mathrm{T}}$$

and $\boldsymbol{\theta}$ is a $q \times k$ matrix.

A parametric quantile function

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \beta(p \mid \boldsymbol{\theta}) = \boldsymbol{x}^T \boldsymbol{\theta} \boldsymbol{b}(p)$$

How to define $\boldsymbol{b}(p)$?

Examples

$$Q(p \mid x, \boldsymbol{\theta}) = \beta_0(p \mid \boldsymbol{\theta}) + \beta_1(p \mid \boldsymbol{\theta})x$$

Example (1)

$$\beta_0(p) = \theta_{00} + \theta_{01} p$$

$$\beta_1(p) = \theta_{10} + \theta_{11} p$$

Coefficients are linear functions of p, defining a Uniform distribution in which both endpoints of the support are linear functions of x

- Different interpretation of the parameters with respect to the "canonical" one.
- Q_1 well defined if $\theta_{01} + \theta_{11} x > 0$ for all x (quantile crossing).
- $\theta_{11} = 0$ forces homoscedasticity.
- $\theta_{00} = \theta_{01} = 0$ (no intercept), a zero-inflated model.

Example (1) - cont.

$$\boldsymbol{b}(p) = \begin{pmatrix} 1 \\ p \end{pmatrix} \text{ and } \boldsymbol{\theta} = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix}$$

Example (2)

$$\beta_0(p) = \theta_{00} + \theta_{01} z(p)$$

$$\beta_1(p) = \theta_{10} + \theta_{12} p$$

, with $z(p)$ the quantile function of a standard Normal.

- A "mix" between Uniform and Normal
- No closed form PDF!
  - This distribution can only be defined through its quantile function.
  - If $\theta_{12} = 0$, reduces to a standard linear model with coefficients $\beta_0 = \theta_{00}$ and $\beta_1 = \theta_{10}$, and residual standard deviation $\sigma = \theta_{01}$.

Example (2) - cont.

$$\boldsymbol{b}(p) = \begin{pmatrix} 1 \\ z(p) \\ p \end{pmatrix} \text{ and } \boldsymbol{\theta} = \begin{pmatrix} \theta_{00} & \theta_{01} & 0 \\ \theta_{10} & 0 & \theta_{12} \end{pmatrix}$$

Example (3)

$$\beta_0(p \mid \boldsymbol{\theta}) = \theta_{00} + \theta_{01} p + \theta_{02} p^2$$

$$\beta_1(p \mid \boldsymbol{\theta}) = \theta_{10} + \theta_{11} \log(p) + \theta_{12} \cos(p)$$

- $\boldsymbol{b}(p)$ must induce a well-defined QF for some $\boldsymbol{\theta}$
- Use meaningful assumptions (e.g., bounded outcome)
- $\boldsymbol{b}(p)$ can have asymptotes

Example (4)

$$\beta_0(p \mid \boldsymbol{\theta}) = \theta_{00} + \theta_{01} \log(p) + \theta_{02} \log(1 - p)$$

$$\beta_1(p \mid \boldsymbol{\theta}) = \theta_{10} + \theta_{13} p$$

- $\beta_0(p)$ unbounded
- $\beta_1(p)$ monotone, bounded between θ_{10} (when $p = 0$) and $\theta_{10} + \theta_{13}$ (when $p = 1$)
- Special cases: Uniform, asymmetric Logistic, Logistic, (shifted) Exponential

The estimator (Frumento and Bottai, 2016)

Ordinary quantile regression for the pth quantile: minimize

$$L_n(\boldsymbol{\beta}(p)) = n^{-1} \sum_{i=1}^{n} (p - \omega_{p,i})(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(p))$$

where $\omega_{p,i} = I(y_i \leq \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(p))$.

Our estimator: minimize

$$\overline{L}_n(\boldsymbol{\theta}) = \int_0^1 L_n(\boldsymbol{\beta}(p \mid \boldsymbol{\theta}))\mathrm{d}p.$$

- Average loss function
- Estimating "all" quantiles at once

The gradient function

Ordinary QR: find the approximated zeros of

$$S_n(\boldsymbol{\beta}(p)) = n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i(\omega_{p,i} - p).$$

Our estimator: find the zeros of

$$\overline{S}_n(\boldsymbol{\theta}) = \int_0^1 S_n(\boldsymbol{\beta}(p \mid \boldsymbol{\theta}))\boldsymbol{b}(p)^{\mathrm{T}} \mathrm{d}p.$$

The estimator (cont.)

We define the following

$$\boldsymbol{B}(p) = \int_0^p \boldsymbol{b}(u)\mathrm{d}u, \;\; \overline{\boldsymbol{B}} = \int_0^1 \boldsymbol{B}(u)\mathrm{d}u$$

This permits writing

$$\overline{L}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n y_i(p_i - 0.5) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}\left[\overline{\boldsymbol{B}} - \boldsymbol{B}(p_i)\right],$$

$$\overline{S}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i \left[\overline{\boldsymbol{B}} - \boldsymbol{B}(p_i)\right]^{\mathrm{T}},$$

where $p_i = F(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ is such that $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(p_i \mid \boldsymbol{\theta}) = y_i$.

Properties

- Smooth objective function
- Simple asymptotic properties
- Could take the integral over (p_1, p_2). Standard QR if $p_1 = p_2 = p$
- More efficient than QR
- $E\left[\overline{S}_n(\boldsymbol{\theta})\right] = 0$ if $p_i \sim U(0,1)$ (testing)
- Other parametrization could be used

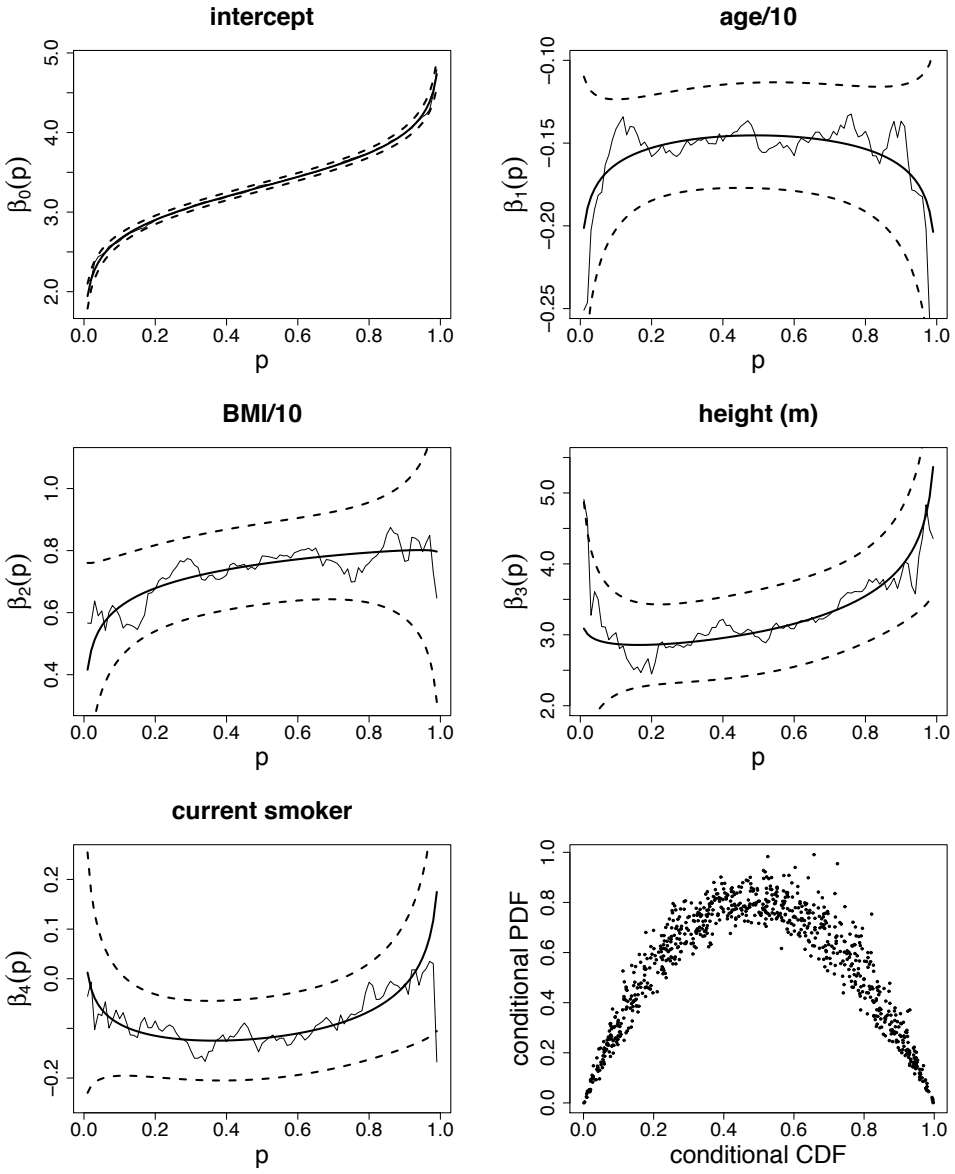
Example

Predicting quantiles of inspiratory capacity based on age, BMI, height, and smoking. For different model specifications, we tested $H_0 : p_i \sim U(0, 1)$.

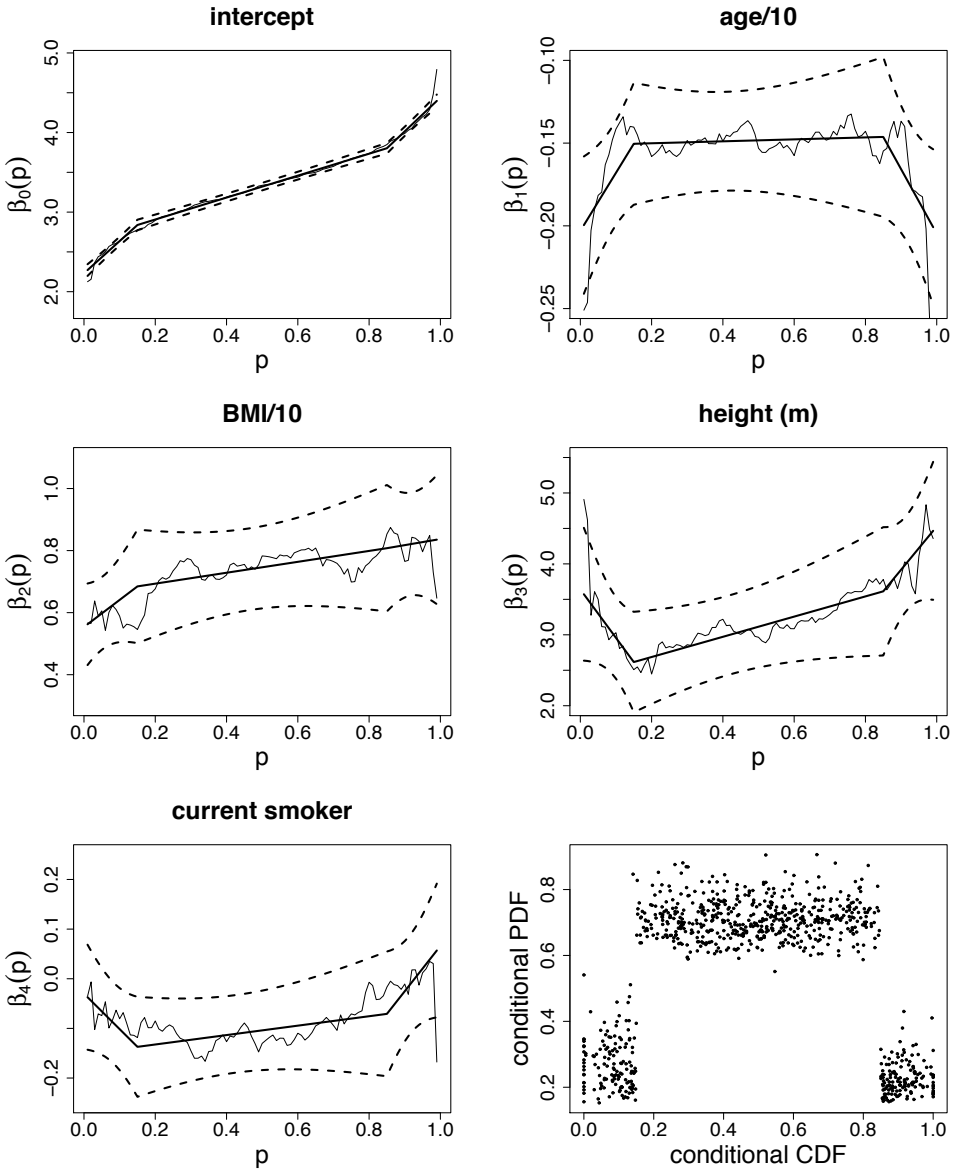
Table: Alternative model specifications

Model	$\boldsymbol{b}(p)$	Loss	P-value H_0
1	p	127.90	0.000
2	p, p^2	127.81	0.000
3	p, p^2, p^3	126.93	0.928
4	$z(p)$ (Normal)	126.98	0.672
5	$\log[p/(1-p)]$	127.01	0.768
6	$\log(p), \log(1-p)$	126.86	0.878
7	piecewise linear	126.98	0.355

Model 6


intercept
$\beta_0(p)$
p
age/10
$\beta_1(p)$
p
BMI/10
$\beta_2(p)$
p
height (m)
$\beta_3(p)$
p
current smoker
$\beta_4(p)$
p
conditional PDF
conditional CDF

Model 7


intercept
$\beta_0(p)$
p
age/10
$\beta_1(p)$
p
BMI/10
$\beta_2(p)$
p
height (m)
$\beta_3(p)$
p
current smoker
$\beta_4(p)$
p
conditional PDF
conditional CDF

Model 7

Table: Summary of model 7

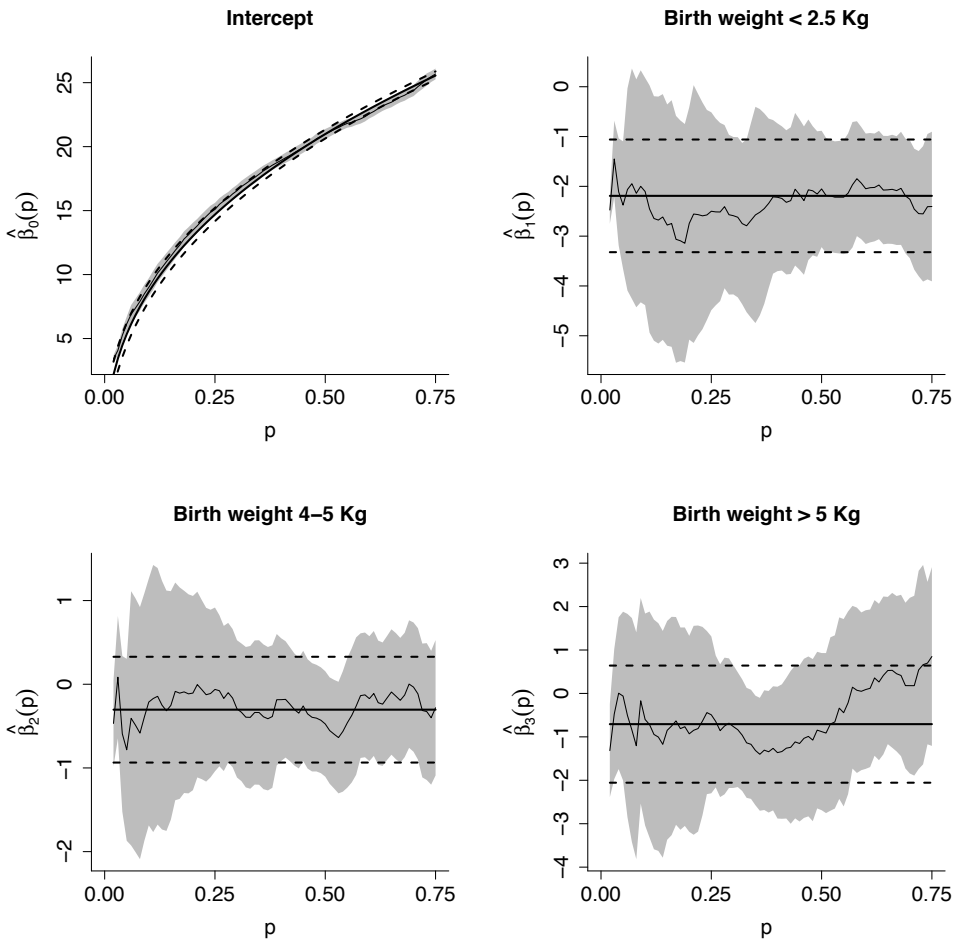
		slope of p			
	Intercept	$p \leq 0.15$	$0.15 < p \leq 0.85$	$p > 0.85$	P-value
Intercept	2.23 (.04)*	4.05 (.32)*	1.38 (.07)*	4.25 (.33)*	0.000*
Age /10	-0.20 (.02)*	0.35 (.17)*	0.01 (.04)	-0.39 (.22)	0.000*
BMI/10	0.55 (.07)*	0.87 (.73)	0.18 (.21)	0.20 (.89)	0.000*
Height (m)	3.64 (.51)*	-6.83 (4.20)	1.42 (.84)	6.10 (4.78)	0.000*
Smoker	-0.03 (.06)	-0.71 (.46)	0.09 (.12)	0.91 (.61)	0.049*
P-value	0.000*	0.000*	0.000*	0.000*	

For each 0.01 increase of p, the quantile regression coefficient associated with BMI/10 increases by 0.87/100 in the range $0 \leq p \leq 0.15$, 0.18/100 in the range $0.15 \leq p < 0.85$, and 0.20/100 in the range $0.85 \leq p \leq 1$

Censored and truncated data

- There is no obvious generalization of $L_n(\beta(p))$.
- However, a generalization of $S_n(\beta(p))$ exists (but this is another story).
- We generalized our estimation equation to the presence of censoring and truncation.

Censored and truncated data: example



Censored and truncated data: example (cont.)

Table: Model summary

	1	$p^{0.4}$
Intercept	-4.99 (0.67)*	34.28 (0.77)*
Birth weight < 2.5	-2.19 (0.58)*	-
Birth weight $4 - 5$	-0.30 (0.32)	-
Birth weight > 5	-0.71 (0.69)	-

Conclusions

- QF modeling is an alternative to PDF (likelihood) modeling
- We developed a new estimator, with specific goodness-of-fit procedures
- Advantages: ease of interpretation, efficiency, parsimony
- Computation... done

References

Frumento, P., and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. Biometrics 72(1), 74-84.

Paolo Frumento (2016). qrcm: Quantile Regression Coefficients Modeling. R package version 1.0 (version 2.0 in progress).