Quantile regression methods and applications: An overview

Marco Geraci

Centre for Paediatric Epidemiology and Biostatistics
UCL Institute of Child Health

26 September 2012

Outline

- Introduction to quantile regression (QR)

Outline

- Introduction to quantile regression (QR)
- Inference

Outline

- Introduction to quantile regression (QR)
- Inference
- Applications

Outline

- Introduction to quantile regression (QR)
- Inference
- Applications
- Software

Outline

- Introduction to quantile regression (QR)
- Inference
- Applications
- Software
- Conclusion

What is it?

Consider the real-valued random variable Y with (right-continuous) distribution function

$$F(y) = P(Y \leq y).$$

The pth quantile of Y, for any $0 < p < 1$, is defined as

$$Q_y(p) \equiv F^{-1}(p) = \inf \{y : F(y) \geq p\}.$$

The pth quantile $\hat{y}$ can be estimated by minimizing the loss function

$$g_p(y - \hat{y}) = \begin{cases} p\,|y - \hat{y}| & \text{if } (y - \hat{y}) \geq 0 \\ (1-p)\,|y - \hat{y}| & \text{if } (y - \hat{y}) < 0 \end{cases}$$

Regression analysis

Consider a predictor X possibly associated with Y. The well-known regression model for the mean

$$\mathrm{E}\,(Y|X = x) = \alpha + \beta x$$

establishes a (linear) relationship between the expectation of Y and $X = x$.

Regression analysis

Consider a predictor X possibly associated with Y. The well-known regression model for the mean

$$\mathrm{E}(Y|X = x) = \alpha + \beta x$$

establishes a (linear) relationship between the expectation of Y and $X = x$.

The pth quantile regression model

$$Q_{Y|X=x}(p) = \alpha(p) + \beta(p)x$$

establishes a (linear) relationship between the pth quantile of Y and $X = x$.

When is the mean sufficient? The location-shift model

- In the *iid* case $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the conditional quantile functions of y

  $$Q_{y|x}(p) = \beta_0 + \beta_1 x + F_{\varepsilon}^{-1}(p) = \alpha(p) + \beta_1 x,$$

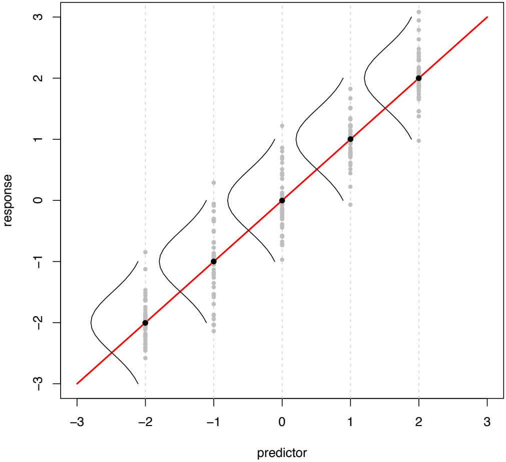
  $\alpha(p) = \beta_0 + F_{\varepsilon}^{-1}(p)$, are simply vertical **translations** of one another. The slope β_1 is the same at all quantile levels.

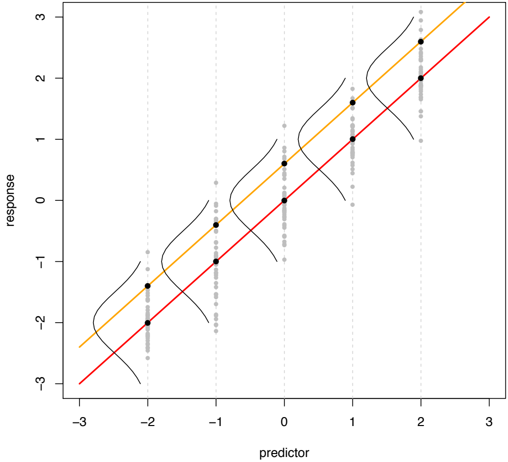
When is the mean sufficient? The location-shift model

- In the *iid* case $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the conditional quantile functions of y

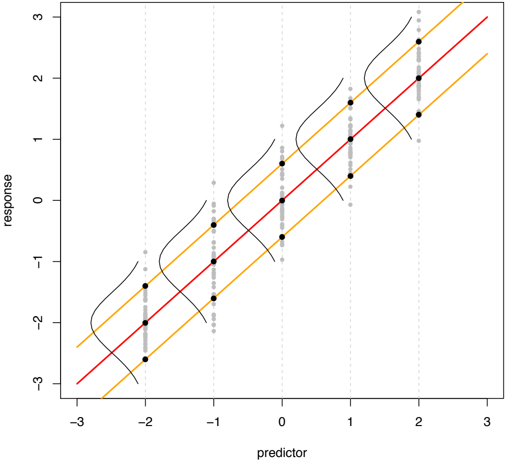
  $$Q_{y|x}(p) = \beta_0 + \beta_1 x + F_\varepsilon^{-1}(p) = \alpha(p) + \beta_1 x,$$

  $\alpha(p) = \beta_0 + F_\varepsilon^{-1}(p)$, are simply vertical **translations** of one another. The slope β_1 is the same at all quantile levels.
- For (approximately) homoscedastic Gaussian errors, the least squares estimation of the parameters is sufficient to make inference about the conditional distribution of y (and with further modelling, heteroscedasticity can be accommodated).


response
predictor
−3
−2
−1
0
1
2
3



response
predictor
−3
−2
−1
0
1
2
3



response
predictor
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

The location–scale-shift model

Consider the *nid* case $y_i = \beta_0 + \beta_1 x_i + (1 + x_i)\varepsilon_i$. The conditional quantile functions of y are

$$Q_{y|x}(p) = \beta_0 + \beta_1 x + F_\varepsilon^{-1}(p) + xF_\varepsilon^{-1}(p) = \alpha(p) + \beta(p)x,$$

$\alpha(p) = \beta_0 + F_\varepsilon^{-1}(p)$ and $\beta(p) = \beta_0 + F_\varepsilon^{-1}(p)$.

The location–scale-shift model

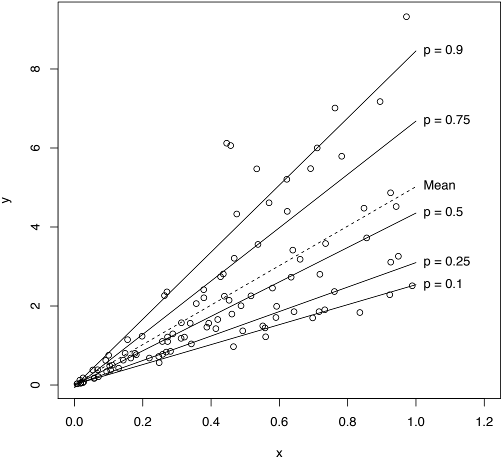
Consider the *nid* case $y_i = \beta_0 + \beta_1 x_i + (1 + x_i)\varepsilon_i$. The conditional quantile functions of y are

$$Q_{y|x}(p) = \beta_0 + \beta_1 x + F_\varepsilon^{-1}(p) + xF_\varepsilon^{-1}(p) = \alpha(p) + \beta(p)x,$$

$\alpha(p) = \beta_0 + F_\varepsilon^{-1}(p)$ and $\beta(p) = \beta_0 + F_\varepsilon^{-1}(p)$.

Let's also add some skewness to ε.

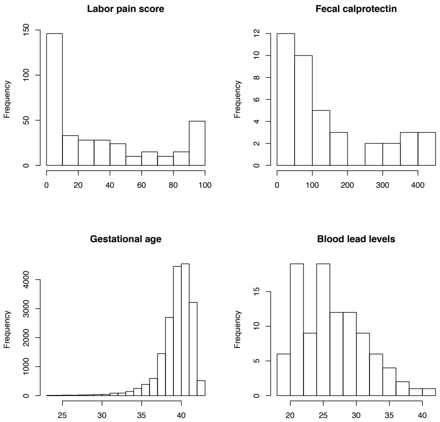
Preview of QR



Comparing mean and median

Should we scare the opposition by announcing our mean height or lull them by announcing our median height?
(Bibby J, *Quotes, Damned Quotes and ..., Edinburgh 1968, p. 21)*

Comparing mean and median (cont'd)



Some empirical marginal distributions

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve “normality” so that inference is made on $\mathrm{E}\{h(Y)\,|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve "normality" so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve “normality” so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve "normality" so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x
- homoscedasticity

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve “normality” so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x
- homoscedasticity
- normality of the error

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve “normality” so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x
- homoscedasticity
- normality of the error

The other challenges are

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve “normality” so that inference is made on $\mathrm{E}\{h(Y)|X=x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x
- homoscedasticity
- normality of the error

The other challenges are

- parameters’ interpretation and inference. We have estimated $\mathrm{E}\{h(Y)|X=x\}$ but usually interest is in $\mathrm{E}\{Y|X=x\}$

To transform or not to transform?

Box-Cox transformations and the like are often used to achieve "normality" so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x
- homoscedasticity
- normality of the error

The other challenges are

- parameters' interpretation and inference. We have estimated $\mathrm{E}\{h(Y)|X = x\}$ but usually interest is in $\mathrm{E}\{Y|X = x\}$
- back-transformation. Recall that in general $\mathrm{E}\{h(Y)|X = x\} \neq h(\mathrm{E}\{Y|X = x)\}$

To transform or not to transform?

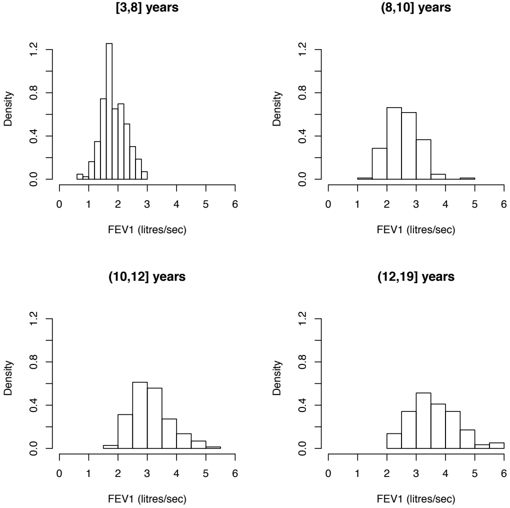
Box-Cox transformations and the like are often used to achieve “normality” so that inference is made on $\mathrm{E}\{h(Y)|X = x\} = \tilde{\alpha} + \tilde{\beta}x$.
This approach presents first a technical challenge as the transformation h should achieve

- linearity in the relationship between mean $h(Y)$ and x
- homoscedasticity
- normality of the error

The other challenges are

- parameters’ interpretation and inference. We have estimated $\mathrm{E}\{h(Y)|X = x\}$ but usually interest is in $\mathrm{E}\{Y|X = x\}$
- back-transformation. Recall that in general $\mathrm{E}\{h(Y)|X = x\} \neq h(\mathrm{E}\{Y|X = x)\}$

In general, the **conditional** distribution of Y might have a complex relationship with X. Location, scale and shape could be all affected by X.



Distribution of FEV1 by age group

Distribution-free QR

Consider n observations of a q-dimensional vector x_i and continuous response y_i.

[1] Koenker and Bassett (1978, *Econometrica*). Earlier work on L_1 norm regression can be traced back to Boscovich (1757), half century before Legendre's work on least squares, and to Edgeworth (1888). See also Wagner (1959, *JASA*).

Distribution-free QR

Consider n observations of a q-dimensional vector x_i and continuous response y_i. The pth linear regression quantile $\hat{\beta}(p)$ is defined as any solution of the asymmetrically weighted absolute deviations problem[1]

$$\min_{\beta} \sum_{r_i \geq 0} p|y_i - x_i'\beta| + \sum_{r_i < 0} (1-p)|y_i - x_i'\beta|,$$

where $r_i = y_i - x_i$.

[1] Koenker and Bassett (1978, *Econometrica*). Earlier work on L_1 norm regression can be traced back to Boscovich (1757), half century before Legendre's work on least squares, and to Edgeworth (1888). See also Wagner (1959, *JASA*).

Distribution-free QR

Consider n observations of a q-dimensional vector x_i and continuous response y_i. The pth linear regression quantile $\hat{\beta}(p)$ is defined as any solution of the asymmetrically weighted absolute deviations problem[1]

$$\min_{\beta} \sum_{r_i \geq 0} p|y_i - x_i'\beta| + \sum_{r_i < 0} (1-p)|y_i - x_i'\beta|,$$

where $r_i = y_i - x_i$.

No assumption on the shape of the error distribution.

[1] Koenker and Bassett (1978, *Econometrica*). Earlier work on L_1 norm regression can be traced back to Boscovich (1757), half century before Legendre's work on least squares, and to Edgeworth (1888). See also Wagner (1959, *JASA*).

Estimation

Linear programming problem

$$\min_{(\beta, r^+, r^-) \in R^q \times R_+^{2n}} \left\{ p 1_n' r^+ + (1-p) 1_n' r^- | X\beta + r^+ - r^- = y \right\}$$

- small to moderate problems ($n \ll 5000$ and $q \ll 20$): Barrodale and Roberts (Koenker and d'Orey, 1987, 1994)
- large: Frisch-Newton interior point method (Portnoy and Koenker, 1997)
- XXL: Frisch-Newton after preprocessing

Standard errors, confidence intervals, and hypothesis testing

An important distinction is made if errors are assumed *iid* or *nid*

Standard errors, confidence intervals, and hypothesis testing

An important distinction is made if errors are assumed *iid* or *nid*

- *iid*: asymptotic covariance matrix (Koenker and Bassett, 1978)

Standard errors, confidence intervals, and hypothesis testing

An important distinction is made if errors are assumed *iid* or *nid*

- *iid*: asymptotic covariance matrix (Koenker and Bassett, 1978)
- *nid*: Huber sandwich estimator based on local estimate of sparsity (i.e., $1/f\left(Q(p)\right)$) and Powell (1991) sandwich estimator based on kernel estimation

Standard errors, confidence intervals, and hypothesis testing

An important distinction is made if errors are assumed *iid* or *nid*

- *iid*: asymptotic covariance matrix (Koenker and Bassett, 1978)
- *nid*: Huber sandwich estimator based on local estimate of sparsity (i.e., $1/f\,(Q(p))$) and Powell (1991) sandwich estimator based on kernel estimation
- *iid* **or** *nid*: rank methods give confidence intervals by inverting a rank test (Koenker, 2005)

Standard errors, confidence intervals, and hypothesis testing

An important distinction is made if errors are assumed *iid* or *nid*

- *iid*: asymptotic covariance matrix (Koenker and Bassett, 1978)
- *nid*: Huber sandwich estimator based on local estimate of sparsity (i.e., $1/f\left(Q(p)\right)$) and Powell (1991) sandwich estimator based on kernel estimation
- *iid* **or** *nid*: rank methods give confidence intervals by inverting a rank test (Koenker, 2005)
- *iid* **and** *nid*: bootstrap (at least five approaches)

Standard errors, confidence intervals, and hypothesis testing

An important distinction is made if errors are assumed *iid* or *nid*

- *iid*: asymptotic covariance matrix (Koenker and Bassett, 1978)
- *nid*: Huber sandwich estimator based on local estimate of sparsity (i.e., $1/f\left(Q(p)\right)$) and Powell (1991) sandwich estimator based on kernel estimation
- *iid* **or** *nid*: rank methods give confidence intervals by inverting a rank test (Koenker, 2005)
- *iid* **and** *nid*: bootstrap (at least five approaches)

Bootstrap performs well in many situations; no need to worry about *iid* or *nid*; computation can be slower than other methods.

Model selection

Purpose of the testing: i) the fitted models all have the same specified quantile and the purpose is to test smaller models vs larger models;

Model selection

Purpose of the testing: i) the fitted models all have the same specified quantile and the purpose is to test smaller models vs larger models; ii) the fitted models have different specified quantile but the same predictor and the purpose is to test if the slopes of the models are identical (e.g., location–shift vs location-scale–shift)

Model selection

Purpose of the testing: i) the fitted models all have the same specified quantile and the purpose is to test smaller models vs larger models; ii) the fitted models have different specified quantile but the same predictor and the purpose is to test if the slopes of the models are identical (e.g., location–shift vs location-scale–shift)

- i) SIC (Machado, 1993), AIC, lasso penalty/shrinkage methods, rank-score tests (Koenker, 2005; Wu and Liu, 2009, *Statistica Sinica*), ANOVA (Chen et al, 2008, *Biometrika*), Wald-type tests (Koenker and Bassett, 1982, *Econometrica*), likelihood ratio tests (Koenker and Bassett, 1982; Koenker and Machado, 1999)

Model selection

Purpose of the testing: i) the fitted models all have the same specified quantile and the purpose is to test smaller models vs larger models; ii) the fitted models have different specified quantile but the same predictor and the purpose is to test if the slopes of the models are identical (e.g., location–shift vs location-scale–shift)

- i) SIC (Machado, 1993), AIC, lasso penalty/shrinkage methods, rank-score tests (Koenker, 2005; Wu and Liu, 2009, *Statistica Sinica*), ANOVA (Chen et al, 2008, *Biometrika*), Wald-type tests (Koenker and Bassett, 1982, *Econometrica*), likelihood ratio tests (Koenker and Bassett, 1982; Koenker and Machado, 1999)
- ii) Khamladze tests, Wald, quantile likelihood ratio and regression rankscore processes (Koenker, 2005)

Advantages

- *depth of analysis*: QR provides a complete picture of the relationships among variables. Location, scale and shape of response distribution conditional on predictors can be investigated

Advantages

- *depth of analysis*: QR provides a complete picture of the relationships among variables. Location, scale and shape of response distribution conditional on predictors can be investigated
- *robustness*: QR is robust to outliers in the outcome

Advantages

- *depth of analysis*: QR provides a complete picture of the relationships among variables. Location, scale and shape of response distribution conditional on predictors can be investigated
- *robustness*: QR is robust to outliers in the outcome
- *efficiency of the median*: as location parameter the median can be more efficient than the mean (e.g., heavy-tailed distributions)

Advantages

- *depth of analysis*: QR provides a complete picture of the relationships among variables. Location, scale and shape of response distribution conditional on predictors can be investigated
- *robustness*: QR is robust to outliers in the outcome
- *efficiency of the median*: as location parameter the median can be more efficient than the mean (e.g., heavy-tailed distributions)
- *equivariance properties*: quantiles are equivariant to monotone transformations

Equivariance properties

Koenker (2005, *Cambridge University Press*)

- $\hat{\beta}(p; ay, X) = a\hat{\beta}(p; y, X)$
- $\hat{\beta}(p; -ay, X) = -a\hat{\beta}(1-p; y, X)$
- $\hat{\beta}(p; y + X\gamma, X) = \hat{\beta}(p; y, X) + \gamma$
- $\hat{\beta}(p; y, XA) = A^{-1}\hat{\beta}(p; y, X)$
- $Q_{h(Y)}(p) = h\left(Q_Y(p)\right)$, where $h(\cdot)$ is a non-decreasing function on $\mathbb{R}$. That is, $\Pr\left(Y \leq y\right) = \Pr\left(h(Y) \leq h(y)\right)$

Caveats

- *uncertainty of estimation in regions where data are less dense*: typically occurs in the tails (...unless distribution is heavy-tailed)

Caveats

- *uncertainty of estimation in regions where data are less dense*: typically occurs in the tails (...unless distribution is heavy-tailed)
- *quantile crossing*: individual (as opposed to joint) estimation of regression quantiles can potentially lead curves to cross. It might be a sign of data sparsity and/or model misspecification. See He (1997, *JASA*) and Zhao (2000, *JMA*) for proposed solution

Caveats

- *uncertainty of estimation in regions where data are less dense*: typically occurs in the tails (...unless distribution is heavy-tailed)
- *quantile crossing*: individual (as opposed to joint) estimation of regression quantiles can potentially lead curves to cross. It might be a sign of data sparsity and/or model misspecification. See He (1997, *JASA*) and Zhao (2000, *JMA*) for proposed solution
- *nondifferentiable objective function*: requires some more sophisticated mathematics (e.g., Bahadur representation, generalized derivatives)

Caveats

- *uncertainty of estimation in regions where data are less dense*: typically occurs in the tails (...unless distribution is heavy-tailed)
- *quantile crossing*: individual (as opposed to joint) estimation of regression quantiles can potentially lead curves to cross. It might be a sign of data sparsity and/or model misspecification. See He (1997, *JASA*) and Zhao (2000, *JMA*) for proposed solution
- *nondifferentiable objective function*: requires some more sophisticated mathematics (e.g., Bahadur representation, generalized derivatives)
- *nonuniqueness of solution*: this is sometimes seen as a downside as compared to least squares optimization; however, multiple solutions can be expressed as linear combinations of basic solution. Also the convexity of the limiting objective function assures the uniqueness of the minimizer (Koenker, 2005)

FEV study

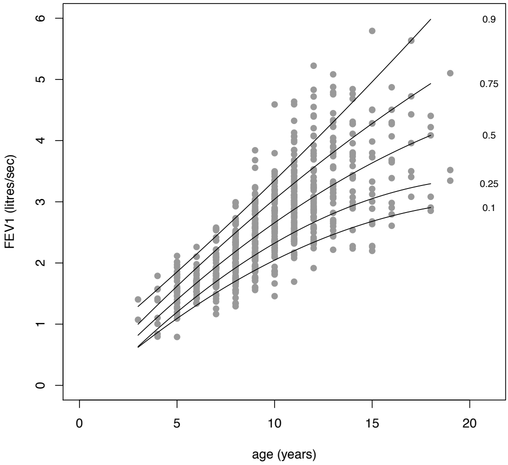
- FEV1 (litres/sec) observed in a sample of 654 individuals aged 3 to 19 years (Childhood Respiratory Disease Study in East Boston, Massachusetts)

FEV study

- FEV1 (litres/sec) observed in a sample of 654 individuals aged 3 to 19 years (Childhood Respiratory Disease Study in East Boston, Massachusetts)
- additional variables: age (years), sex, height (inches) and smoking status

FEV study

- FEV1 (litres/sec) observed in a sample of 654 individuals aged 3 to 19 years (Childhood Respiratory Disease Study in East Boston, Massachusetts)
- additional variables: age (years), sex, height (inches) and smoking status
- $Q_{\mathrm{FEV}|\mathrm{age}}(p) = \beta_0(p) + \beta_1(p)\mathrm{age} + \beta_2(p)\mathrm{age}^2$



Quantile regression curves for FEV vs age

Quantile-based statistics

- *Central tendency*. The median $Q(0.5)$ is a robust alternative to the mean and, depending on the distribution, can be more efficient.

Quantile-based statistics

- *Central tendency*. The median $Q(0.5)$ is a robust alternative to the mean and, depending on the distribution, can be more efficient.
- *Dispersion*. The interquantile range $Q(1-p) - Q(p)$, $p \neq 0.5$ (e.g., $\mathrm{IQR} = Q(0.75) - Q(0.25)$) is a robust alternative to the variance.

Quantile-based statistics

- *Central tendency*. The median $Q(0.5)$ is a robust alternative to the mean and, depending on the distribution, can be more efficient.
- *Dispersion*. The interquantile range $Q(1-p) - Q(p)$, $p \neq 0.5$ (e.g., $\mathrm{IQR} = Q(0.75) - Q(0.25)$) is a robust alternative to the variance.
- *Shape*. The asymmetry index

$$A_p = \frac{Q(1-p) - Q(0.5)}{Q(0.5) - Q(p)},$$

for $p < 0.5$ and $Q(p) < Q(0.5)$,

Quantile-based statistics

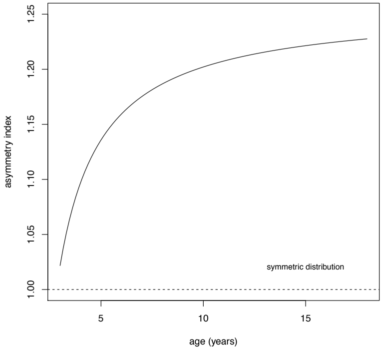
- *Central tendency*. The median $Q(0.5)$ is a robust alternative to the mean and, depending on the distribution, can be more efficient.
- *Dispersion*. The interquantile range $Q(1-p) - Q(p)$, $p \neq 0.5$ (e.g., $\mathrm{IQR} = Q(0.75) - Q(0.25)$) is a robust alternative to the variance.
- *Shape*. The asymmetry index

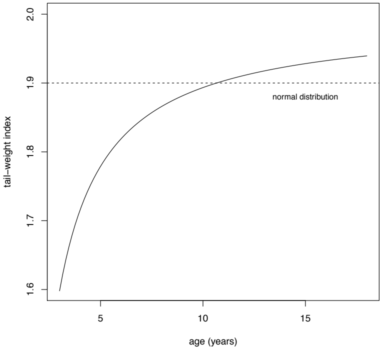
$$A_p = \frac{Q\,(1-p) - Q\,(0.5)}{Q\,(0.5) - Q\,(p)},$$

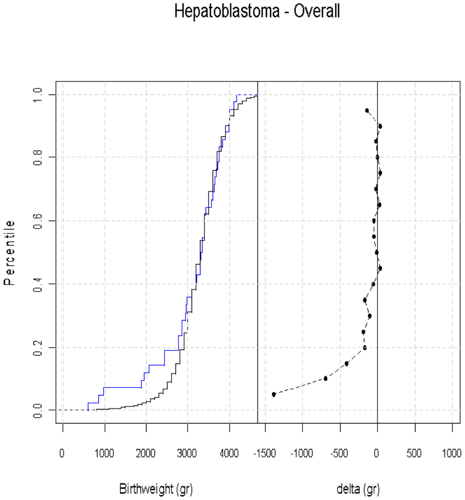
for $p < 0.5$ and $Q\,(p) < Q\,(0.5)$, and the tail-weight index

$$T_p = \frac{Q\,(1-p) - Q\,(p)}{Q\,(0.75) - Q\,(0.25)} = \frac{Q\,(1-p) - Q\,(p)}{\mathrm{IQR}},$$

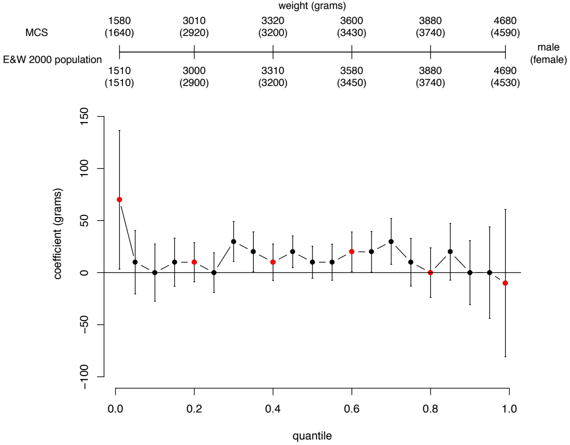
for $p < 0.25$ and $Q\,(0.25) < Q\,(0.75)$.


asymmetry index
1.00
1.05
1.10
1.15
1.20
1.25
symmetric distribution
5
10
15
age (years)


tail–weight index
1.6
1.7
1.8
1.9
2.0
normal distribution
5
10
15
age (years)



Comparison of birthweight distributions: children with hepatoblastoma and the general population $Q(p) = \alpha(p) + \delta(p)I$ (Birch et al, 2010, SIOP, Boston)



Comparison of birthweight distributions: Millennium Cohort Study and the general population $Q(p) = \alpha(p) + \delta(p)I$ (Geraci et al, 2012, MRC Population and Health, Birmingham)

The many uses of the equivariance property

Recall that $Q_{h(Y)|X=x}(p) = h\left(Q_{Y|X=x}(p)\right)$ for a non-decreasing transformation. For example, consider $h(\cdot) \equiv \log(\cdot)$, $Q_{\log(Y)|X=x}(p) = \tilde{\alpha}(p) + \tilde{\beta}(p)x$. Then

$$\frac{\partial Q_{Y|X=x}(p)}{\partial x} = \frac{\partial h^{-1}\left\{Q_{h(Y)|X=x}(p)\right\}}{\partial x} = \exp\left\{\tilde{\alpha}(p) + \tilde{\beta}(p)x\right\}\beta(p)$$

Quantile regression for counts

Consider a random variable Y resulting from a counting process. Define the variable $Z = Y + U$, where U is uniform noise (jittering) with $U \perp Y, X$. Assume the model (Machado and Santos Silva, 2005, *JASA*)

$$Q_{h(Z;p)}(p) = x'\beta(p),$$

where the transformation h possibly depends on p. For example,

$$h(Z;p) = \begin{cases} \log(Z - p) & \text{for } Z > p \\ \log(\delta) & \text{for } Z \leq p \end{cases}$$

with $\delta > 0$ suitably small.

Survival analysis

Many survival models can be expressed as transformation models (Koenker, 2005)

$$h(T_i) = x_i'\beta + \varepsilon_i$$

$h(\cdot)$ monotone and $\varepsilon_i \overset{iid}{\sim} F$.

Survival analysis

Many survival models can be expressed as transformation models (Koenker, 2005)

$$h\,(T_i) = x_i'\beta + \varepsilon_i$$

$h(\cdot)$ monotone and $\varepsilon_i \overset{iid}{\sim} F$.

- Cox proportional hazard model: $\log \Lambda_0\,(T_i) = x_i'\beta + \varepsilon_i$, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ and $F_\varepsilon(u) = 1 - e^{e^u}$

Survival analysis

Many survival models can be expressed as transformation models (Koenker, 2005)

$$h(T_i) = x_i'\beta + \varepsilon_i$$

$h(\cdot)$ monotone and $\varepsilon_i \overset{iid}{\sim} F$.

- Cox proportional hazard model: $\log \Lambda_0(T_i) = x_i'\beta + \varepsilon_i$, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ and $F_\varepsilon(u) = 1 - e^{e^u}$
- AFT model: $\log(T_i) = x_i'\beta + \varepsilon_i$, $\Pr(T > t) = 1 - F\left(te^{-x'\beta}\right)$

Survival analysis

Many survival models can be expressed as transformation models (Koenker, 2005)

$$h\left(T_i\right) = x_i'\beta + \varepsilon_i$$

$h(\cdot)$ monotone and $\varepsilon_i \overset{iid}{\sim} F$.

- Cox proportional hazard model: $\log \Lambda_0\left(T_i\right) = x_i'\beta + \varepsilon_i$, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ and $F_\varepsilon(u) = 1 - e^{e^u}$
- AFT model: $\log\left(T_i\right) = x_i'\beta + \varepsilon_i$, $\Pr\left(T > t\right) = 1 - F\left(te^{-x'\beta}\right)$

The QR model $Q_{h(T)} = x'\beta$ abandons the *iid* assumptions and different forms of heterogeneity in the distribution of $h(T)$ can be introduced. See for example Koenker and Geling (2001, *JASA*).

Censored quantile regression

Consider the latent variable model

$$T_i = x_i'\beta + \varepsilon_i,$$

$\varepsilon_i \sim F$ (or any non-decreasing transformation $h(T)$). If censoring times C_i are observed for all i and $Z_i = \min(C_i, T_i)$ then we can estimate (Powell, 1984, 1986, *J Econ*)

$$\sum g_p \left\{ Z_i - \min\left\{ C_i, x'\beta(p) \right\} \right\}.$$

Binary data

Suppose we observe $Z = I(Y^* > 0)$ and Y^* is a latent process. Suppose also $Q_{Y^*}(p) = x'\beta(p)$. Therefore $Q_{h(Y^*)}(p) = Q_Z(p)$ with $h(Y) = I(Y^* > 0)$. We can estimate $\beta(p)$ by minimizing

$$\sum g_p \left\{ z - I(x'\beta(p) > 0) \right\},$$

where g_p is the QR loss function. This can be also seen as extreme censoring.

For computational issues and advantages of this approach, see Koenker (2005).

Bounded outcomes

Consider a continuous variable y bounded in $[a, b] \subseteq \mathbb{R}$ (e.g., a clinical score). Let's apply (Bottai et al, 2009, *Stats in Med*)

$$h(y) = \log\left(\frac{z_y}{1 - z_y}\right)$$

where $z_y = \frac{y-(a-\delta)}{(b+\delta)-(a-\delta)}$ and $\delta > 0$ is a suitably small quantity so that $0 < z_y < 1$. Then

$$h^{-1}\left(Q_{h(Y)}\right) = \hat{Q}_Y$$

will be bounded in $[a, b]$.

Multiple imputation

In multiple imputation by chained equations, the mean regression model is typically the default choice for continuous variables. However, if the location-shift hypothesis does not hold, QR-based imputation can be used (Bottai and Zhen, 2013; Geraci, 2012, under review).

Multiple imputation

In multiple imputation by chained equations, the mean regression model is typically the default choice for continuous variables. However, if the location-shift hypothesis does not hold, QR-based imputation can be used (Bottai and Zhen, 2013; Geraci, 2012, under review).Recall that if $V \sim F_v$ and $U \sim \mathrm{Unif}(0, 1)$, then $F_v^{-1}(U)$ is distributed as V (probability integral transform).

Multiple imputation

In multiple imputation by chained equations, the mean regression model is typically the default choice for continuous variables. However, if the location-shift hypothesis does not hold, QR-based imputation can be used (Bottai and Zhen, 2013; Geraci, 2012, under review).Recall that if $V \sim F_v$ and $U \sim \mathrm{Unif}(0, 1)$, then $F_v^{-1}(U)$ is distributed as V (probability integral transform). Let $X_{(j)}$ be X without x_j. Take a sample $u \sim \mathrm{Unif}(0, 1)$ and estimate (complete-data only)

$$Q_{x_j|X_{(j)}}(u) = g\left(X_{(j)}\right)$$

for some function g (linear or nonlinear).

Multiple imputation

In multiple imputation by chained equations, the mean regression model is typically the default choice for continuous variables. However, if the location-shift hypothesis does not hold, QR-based imputation can be used (Bottai and Zhen, 2013; Geraci, 2012, under review).Recall that if $V \sim F_v$ and $U \sim \mathrm{Unif}(0,1)$, then $F_v^{-1}(U)$ is distributed as V (probability integral transform). Let $X_{(j)}$ be X without x_j. Take a sample $u \sim \mathrm{Unif}(0,1)$ and estimate (complete-data only)

$$Q_{x_j|X_{(j)}}(u) = g\left(X_{(j)}\right)$$

for some function g (linear or nonlinear). Then impute $\hat{Q}_{x_j|X_{(j)}}(u)$ where x_j is missing.

Multiple imputation

In multiple imputation by chained equations, the mean regression model is typically the default choice for continuous variables. However, if the location-shift hypothesis does not hold, QR-based imputation can be used (Bottai and Zhen, 2013; Geraci, 2012, under review).Recall that if $V \sim F_v$ and $U \sim \mathrm{Unif}(0,1)$, then $F_v^{-1}(U)$ is distributed as V (probability integral transform). Let $X_{(j)}$ be X without x_j. Take a sample $u \sim \mathrm{Unif}(0,1)$ and estimate (complete-data only)

$$Q_{x_j|X_{(j)}}(u) = g\left(X_{(j)}\right)$$

for some function g (linear or nonlinear). Then impute $\hat{Q}_{x_j|X_{(j)}}(u)$ where x_j is missing. Repeat m times and summarize using Rubin's rules.

Multiple imputation

In multiple imputation by chained equations, the mean regression model is typically the default choice for continuous variables. However, if the location-shift hypothesis does not hold, QR-based imputation can be used (Bottai and Zhen, 2013; Geraci, 2012, under review).Recall that if $V \sim F_v$ and $U \sim \mathrm{Unif}(0, 1)$, then $F_v^{-1}(U)$ is distributed as V (probability integral transform). Let $X_{(j)}$ be X without x_j. Take a sample $u \sim \mathrm{Unif}(0, 1)$ and estimate (complete-data only)

$$Q_{x_j|X_{(j)}}(u) = g\left(X_{(j)}\right)$$

for some function g (linear or nonlinear). Then impute $\hat{Q}_{x_j|X_{(j)}}(u)$ where x_j is missing. Repeat m times and summarize using Rubin's rules. QR imputation

- may increase efficiency and reduces bias
- handles a variety of conditional distributions
- handles bounded outcomes (e.g., gestational age, birthweights, etc)

Transformations to linearity

Consider the nonlinear QR model

$$Q_{Y|X}(p) = g\left(x'(p)\beta(p), \lambda_p\right)$$

where g is strictly monotonically increasing in $x'(p)\beta(p)$ and λ_p is a 'transformation' parameter to be estimated.

Transformations to linearity

Consider the nonlinear QR model

$$Q_{Y|X}(p) = g\left(x'(p)\beta(p), \lambda_p\right)$$

where g is strictly monotonically increasing in $x'(p)\beta(p)$ and λ_p is a 'transformation' parameter to be estimated.

Therefore

$$Q_{Y_{\lambda_p}} = x'(p)\beta(p)$$

where $Y_\lambda = g^{-1}(Y, \lambda)$.

Transformations to linearity

Consider the nonlinear QR model

$$Q_{Y|X}(p) = g\left(x'(p)\beta(p), \lambda_p\right)$$

where g is strictly monotonically increasing in $x'(p)\beta(p)$ and λ_p is a 'transformation' parameter to be estimated.

Therefore

$$Q_{Y_{\lambda_p}} = x'(p)\beta(p)$$

where $Y_\lambda = g^{-1}(Y, \lambda)$.

For Box-Cox transformations see Powell (1991), Chamberlain (1994, *Cambridge University Press*), Buchinsky (1995, *J Econometrics*), Fitzenberger et al (2010, *Econ Rev*). Also Aranda-Ordaz transformations for bounded outcomes (Dehbi, Corina-Borja, Geraci, in preparation).

And then came the asymmetric Laplace

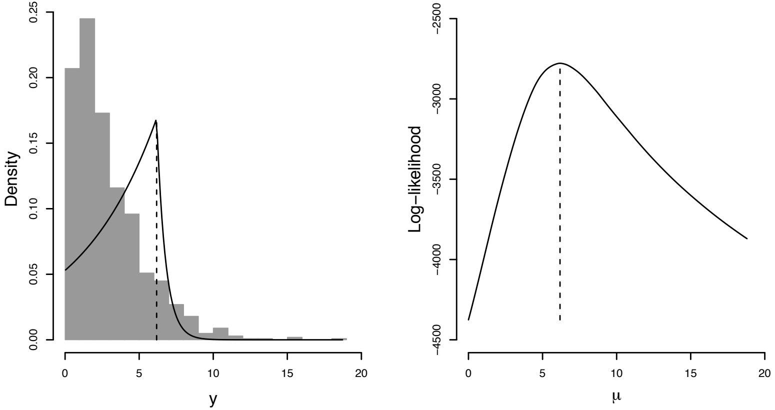
Consider the following density

$$f(y;\mu,\sigma,p) = \frac{p(1-p)}{\sigma}\exp\left\{-\frac{1}{\sigma}g_p\left(y-\mu\right)\right\}$$

where

$$g_p\left(y-\mu\right) = \begin{cases} p\left|y-\mu\right| & \text{if } (y-\mu) \geq 0 \\ (1-p)\left|y-\mu\right| & \text{if } (y-\mu) < 0 \end{cases}$$

and $\Pr(y \leq \mu) = p$ (that is, μ is the p*th* quantile of y).



Maximum likelihood estimation of sample quantiles using the asymmetric Laplace distribution

Likelihoods and loss functions

Mean regression problem
(least squares)

$$\min_{\beta} \sum \left(y - x'\beta\right)^2$$

$$\updownarrow$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y - x'\beta\right)^2\right\}$$

Normal distribution

Quantile regression problem
(least absolute deviations)

$$\min_{\beta} \sum g_p\left(y - x'\beta\right)$$

$$\updownarrow$$

$$\frac{p(1-p)}{\sigma} \exp\left\{-\frac{1}{\sigma} g_p\left(y - x'\beta\right)\right\}$$

Asymmetric Laplace distribution

Likelihoods and loss functions

Mean regression problem
(least squares)

$$\min_{\beta} \sum (y - x'\beta)^2$$

Quantile regression problem
(least absolute deviations)

$$\min_{\beta} \sum g_p (y - x'\beta)$$

location parameter

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y - x'\beta)^2\right\}$$

Normal distribution

$$\frac{p(1-p)}{\sigma} \exp\left\{-\frac{1}{\sigma} g_p (y - x'\beta)\right\}$$

Asymmetric Laplace distribution

Likelihoods and loss functions

Mean regression problem
(least squares)

$$\min_{\beta} \sum \left(y - x'\beta\right)^2$$

Quantile regression problem
(least absolute deviations)

$$\min_{\beta} \sum g_p\left(y - x'\beta\right)$$

scale parameter

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y - x'\beta\right)^2\right\}$$

Normal distribution

$$\frac{p(1-p)}{\sigma} \exp\left\{-\frac{1}{\sigma} g_p\left(y - x'\beta\right)\right\}$$

Asymmetric Laplace distribution

Likelihoods and loss functions

Mean regression problem
(least squares)

$$\min_{\beta} \sum (y - x'\beta)^2$$

Quantile regression problem
(least absolute deviations)

$$\min_{\beta} \sum g_p(y - x'\beta)$$

skewness parameter

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y - x'\beta)^2\right\}$$

Normal distribution

$$\frac{p(1-p)}{\sigma} \exp\left\{-\frac{1}{\sigma} g_p(y - x'\beta)\right\}$$

Asymmetric Laplace distribution

Laplace estimation

- fast computation

Laplace estimation

- fast computation
- Bayesian modelling

Laplace estimation

- fast computation
- Bayesian modelling
- model extensions (e.g., inclusion random effects)

Clustered data

Consider clustered data in the form $(y_{ij}, x'_{ij}, z'_{ij})$. The pth linear quantile mixed model (Geraci and Bottai, 2007, *Biostatistics*; Geraci and Bottai, 2012, under review) is defined as

$$y_{ij} = x'_{ij}\beta(p) + z'_{ij}u + \varepsilon_{ij}(p)$$

where $u \sim G_{\theta_p}$ is a vector of random effects and $\varepsilon_{ij}(p) \sim AL\,(0, \sigma, p)$.

Clustered data

Consider clustered data in the form $(y_{ij}, x'_{ij}, z'_{ij})$. The pth linear quantile mixed model (Geraci and Bottai, 2007, *Biostatistics*; Geraci and Bottai, 2012, under review) is defined as

$$y_{ij} = x'_{ij}\beta(p) + z'_{ij}u + \varepsilon_{ij}(p)$$

where $u \sim G_{\theta_p}$ is a vector of random effects and $\varepsilon_{ij}(p) \sim AL\,(0, \sigma, p)$.

Koenker (2004, *JMA*) considers

$$Q_{Y_i|X=x_i}(p) = x'_i\beta(p) + \alpha_i$$

with L_1 penalized individual fixed intercepts α_i.

Clustered data

Consider clustered data in the form $(y_{ij}, x'_{ij}, z'_{ij})$. The pth linear quantile mixed model (Geraci and Bottai, 2007, *Biostatistics*; Geraci and Bottai, 2012, under review) is defined as

$$y_{ij} = x'_{ij}\beta(p) + z'_{ij}u + \varepsilon_{ij}(p)$$

where $u \sim G_{\theta_p}$ is a vector of random effects and $\varepsilon_{ij}(p) \sim AL\,(0, \sigma, p)$.

Koenker (2004, *JMA*) considers

$$Q_{Y_i|X=x_i}(p) = x'_i\beta(p) + \alpha_i$$

with L_1 penalized individual fixed intercepts α_i.

See Jung (1996, *JASA*) and Lipsitz et al (2000, *JRSS-C*) for estimating equations approach; Yuan and Yin (2010, *Biometrics*) and Reich et al (2010, *Biostatistics*), Bayesian approach; Farcomeni (2012, *Stat and Comp*), latent Markov models.

Spatial quantile regression

Consider a model of the type

$$Q_{Y|\mathbf{x}} = g(\mathbf{x})$$

where g is a linear or nonlinear function of a vector of geographical coordinates $\mathbf{x}$.

Spatial quantile regression

Consider a model of the type

$$Q_{Y|\mathbf{x}} = g(\mathbf{x})$$

where g is a linear or nonlinear function of a vector of geographical coordinates $\mathbf{x}$.Approaches

- bivariate L_1-norm smoothing (triogram splines) (Koenker and Mizera, 2004, *Stat Methodology*)
- Bayesian asymmetric Laplace model for counts using Machado and Santos Silva's (2005) results (Lee and Neocleous, 2010, JRSS-C)
- Bayesian asymmetric Laplace process (Lum and Gelfand, 2012)
- Bayesian infinite mixtures (Reich et al, 2010, *JASA*)

Type of analysis and software

- *response*: univariate cont.[1,2,3,4], multivariate cont., counts[2], binary, clustered[2] (multiple random effects), clustered[1] (fixed intercepts), autocorrelated[1]
- *predictor*: linear[1,2,3,4], non-linear[1], non-parametric[1] (univariate & bivariate splines)
- *analysis*: frequentist[1,2,3,4], Bayesian[5], survival, censored[1,3], spatial (frequentist and Bayesian), complex surveys, missing data, Box-Cox, Aranda-Ordaz[6]

[1] R package `quantreg` (R. Koenker), [2] R package `lqmm` (M. Geraci), [3] Stata, [4] SAS, [5] R package `bayesQR`, [6] R package `AOfamilies` (H. Dehbi)

Conclusion

- QR is a flexible and robust approach in a number of practical situations
- basic theory is well-developed and new advances are progressively being made
- it can be used as a valuable exploratory tool
- software is widely available and easy to use