Quantile Regression

A gentle introduction from scratch to self–sufficiency

Domenico Vistocco

Pisa, 14 July 2016

Department of Economics and Law
University of Cassino and Southern Lazio, Italy

La carte

Motivation

The mean, the whole mean and nothing but the mean

The top 10 reason to become a statistician:

1. Deviation is considered normal
2. We feel complete and sufficient
3. We are “mean” lovers
4. Statistician do it discretely and continuously
5. We are right 95% of the time
6. We can legally comment on someone’s posterior distribution
7. We may not be normal but we are transformable
8. We never have to say we are certain
9. We are honestly significant different
10. No one wants our jobs

Beyond the mean

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of X's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set.

Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

Mostseller and Tukey (1977)

Regression models

Let us consider data $(y_i, \mathbf{z}_i)$ for a continuous response variable y and a set of covariates $\mathbf{Z}$.

The typical regression model is:

$$y_i = \eta_i + \epsilon_i$$

where η_i is a regression predictor formed in terms of the covariates $\mathbf{z}_i$.

NOTE: we will restrict to the case of linear effects:

$$\eta = \beta_0 + \mathbf{x}^\top \beta$$

Regression models

Minimal assumptions on the error term:

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$
- $Cov(\epsilon_i, \epsilon_j) = 0$

Typical assumptions on the error term:

- $\epsilon_i \sim N(0, \sigma^2)$
- independence of ϵ_i and ϵ_j

Regression models

From the assumption on the error term, we have the following properties of the response distribution:

- the predictor η_i determines the expectation of the response:

$$E(y_i|\mathbf{z}_i) = \eta_i$$

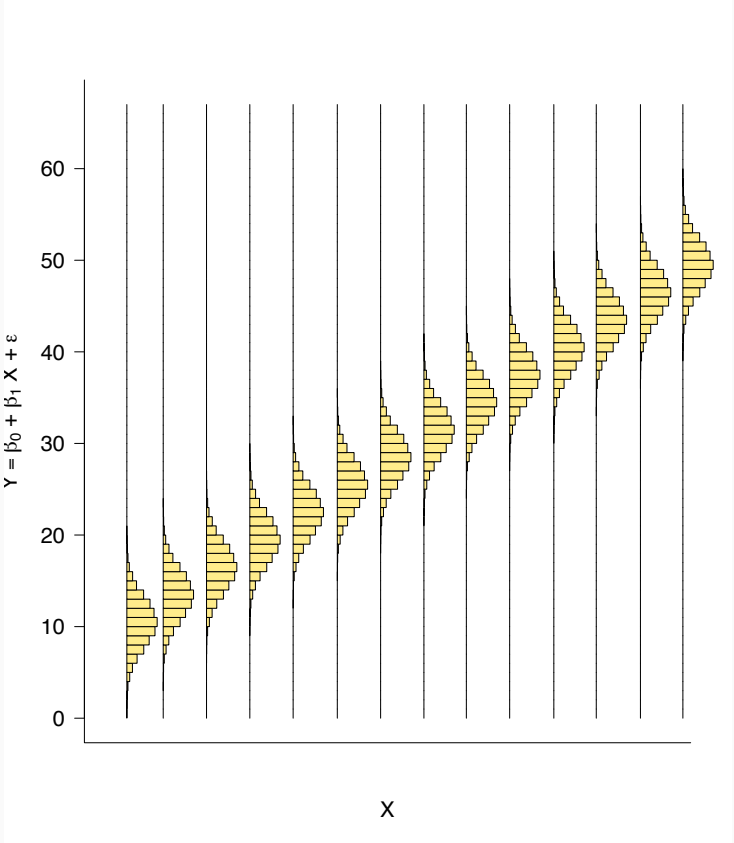
- the response is homoschedastic

$$Var(y_i|\mathbf{z}_i) = \sigma^2$$

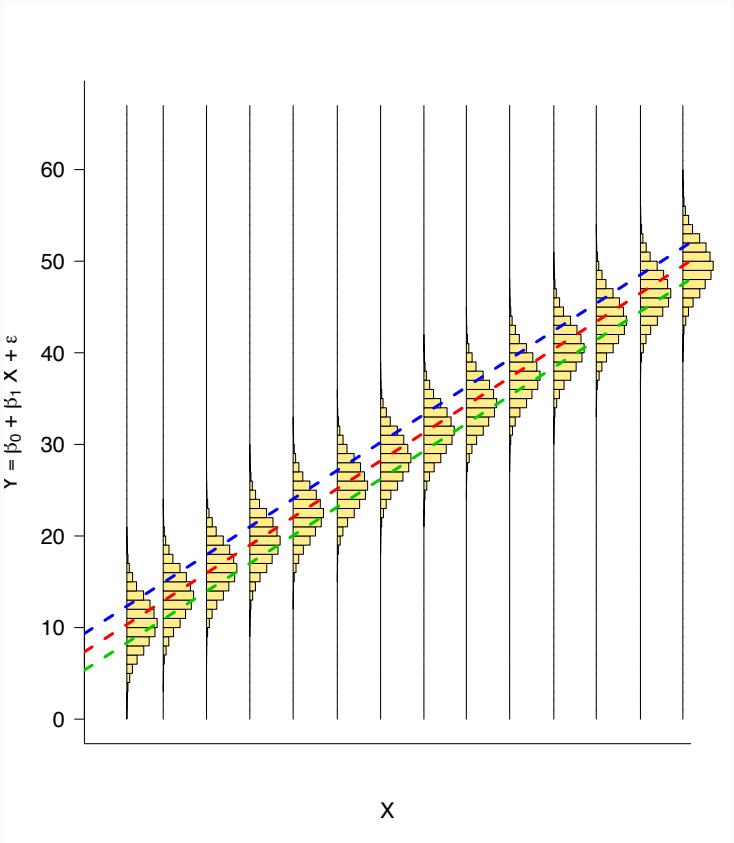
- the quantile curves are parallel
  - in case of normal erros:

$$Q(y_i|\mathbf{z}_i) = \eta_i + z_\tau \sigma$$

The case of normal errors



The case of normal errors

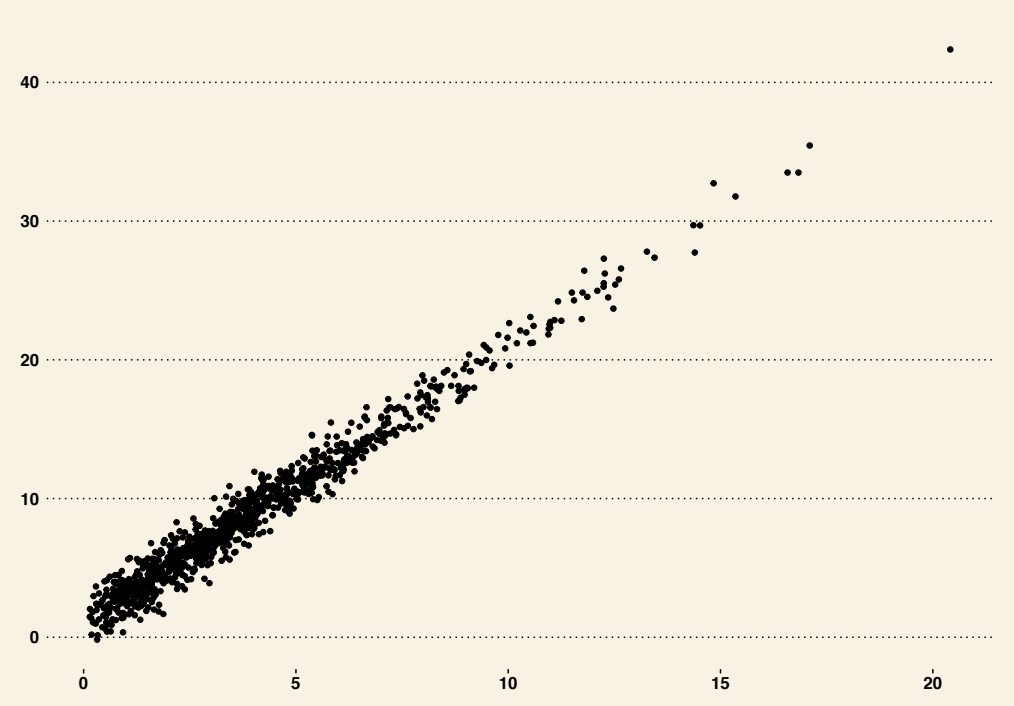


Location-shift model

The case of normal errors

```r
# a normal error (classic) model
set.seed(17)
n <- 1000
df_chi <- 2; beta0 <- 1; beta1 <- 2
x <- rchisq(n, df_chi)
error <- rnorm(n)
y <- beta0 + beta1 * x + error
# organize data into a dataframe
df <- data.frame(x, y)
# scatter plot
library(ggplot2); library(ggthemes)
g1 <- ggplot(data = df, aes(x, y)) +
  geom_point() + theme_wsj()
```


The case of normal errors

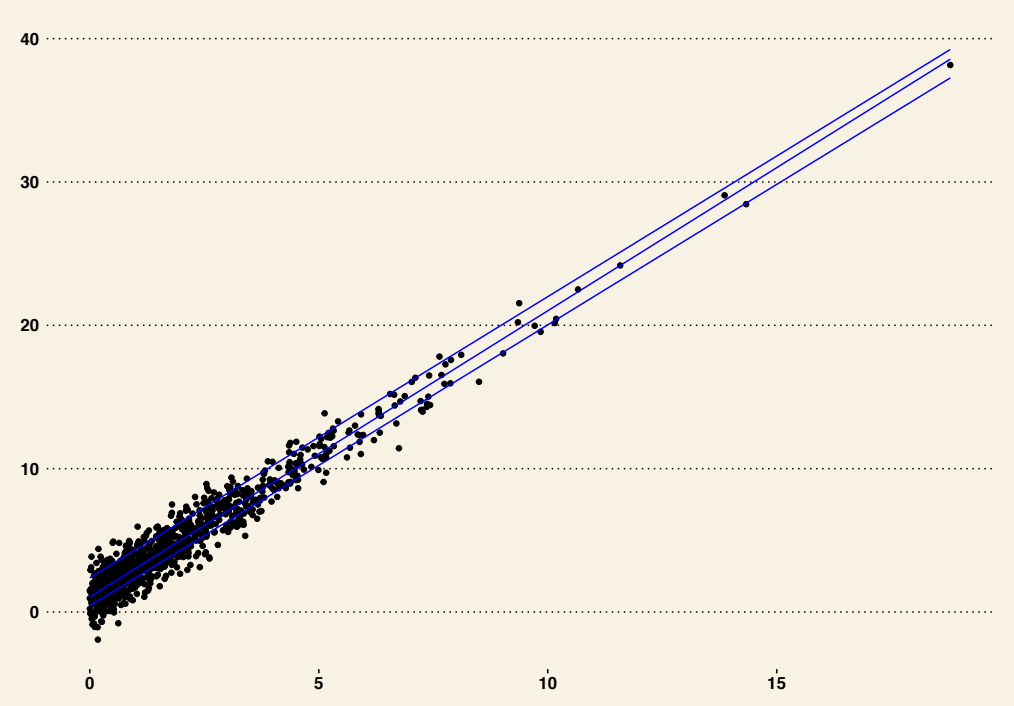


The case of normal errors

```
lm_coefs <- coef(lm(y ~ x, data = df))

g2 <- g1 +
      geom_abline(intercept = lm_coefs[1],
                  slope = lm_coefs[2],
                  colour = "red",
                  linetype = "dashed", size = 2) +
      stat_quantile(quantiles = c(0.25, 0.5, 0.9),
                    colour = "blue")
```

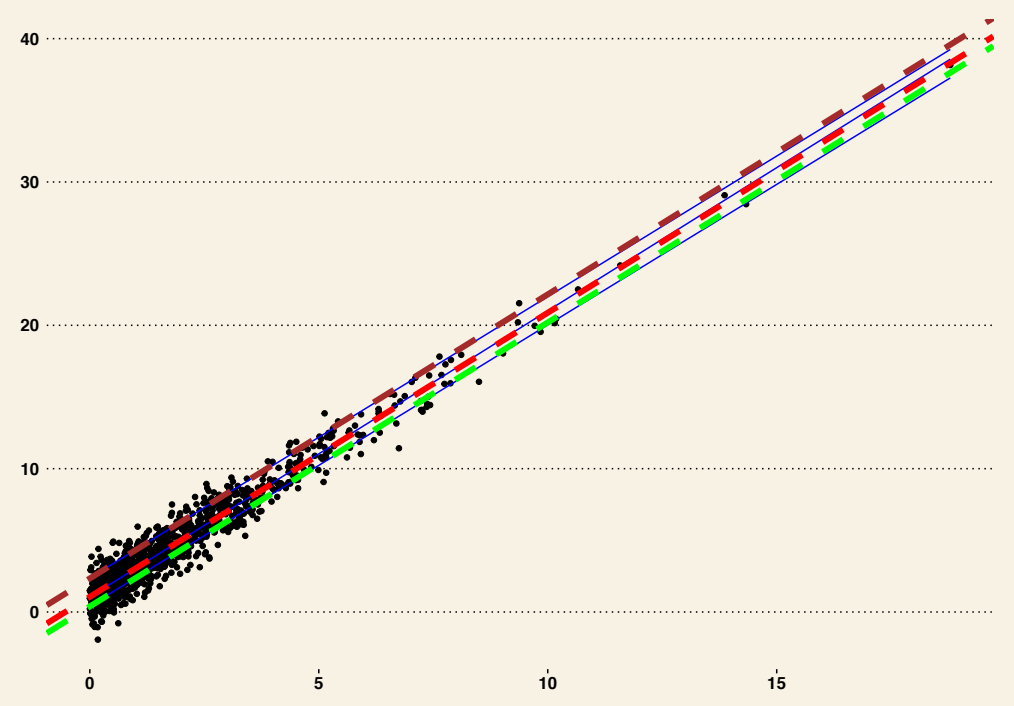

The case of normal errors



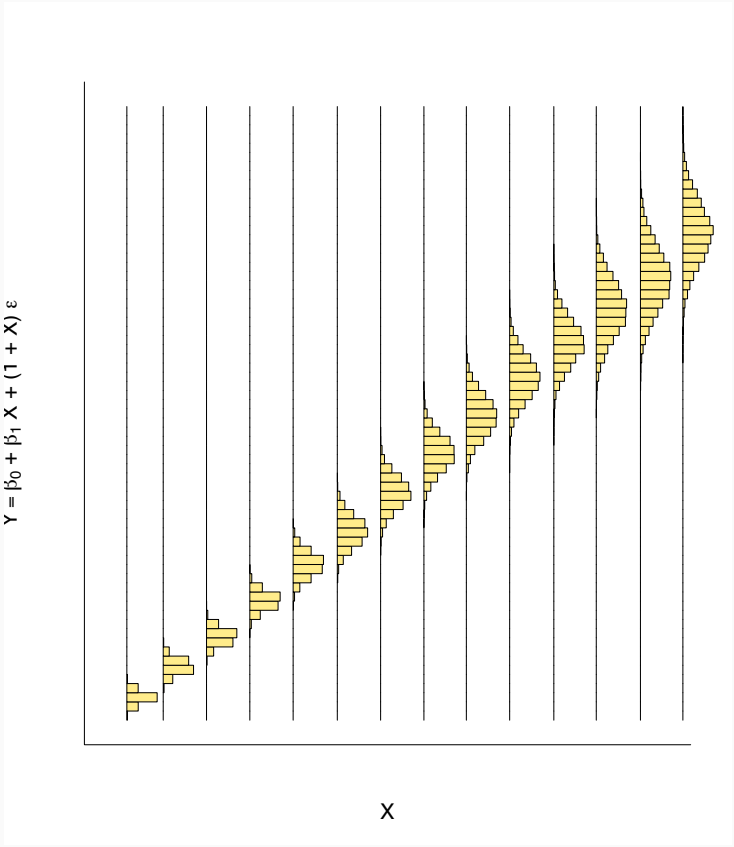
The case of normal errors

```
g3 <- g2 +
  geom_abline(intercept = lm_coefs[1] +
                qnorm(0.25) *
                summary(lm(y ~ x, data = df))$sigma,
              slope = lm_coefs[2],
              colour = "green",
              linetype = "dashed", size = 2) +
  geom_abline(intercept = lm_coefs[1] + qnorm(0.9) *
                summary(lm(y ~ x, data = df))$sigma,,
              slope = lm_coefs[2],
              colour = "brown",
              linetype = "dashed", size = 2)
```

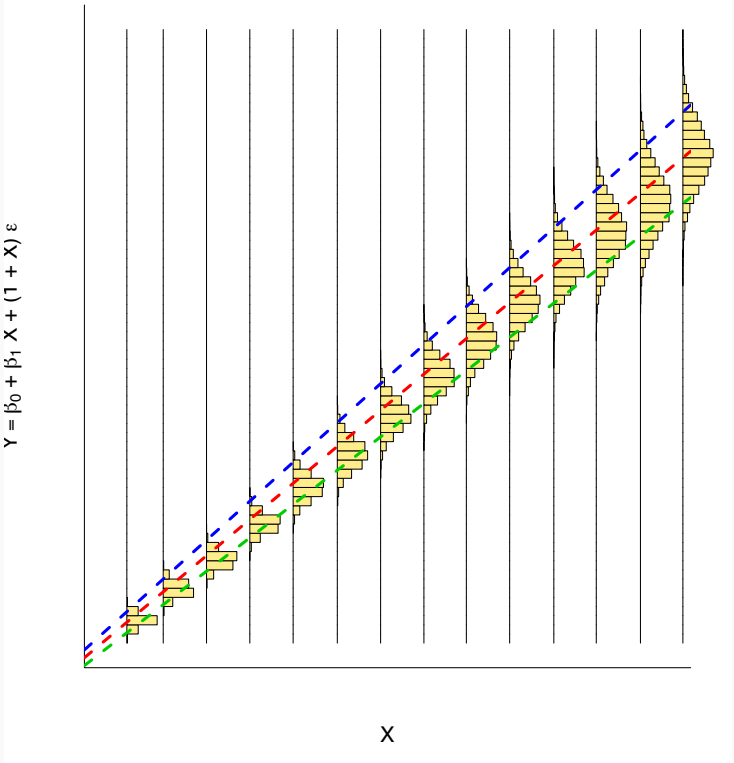

The case of normal errors



The case of heterogeneous errors

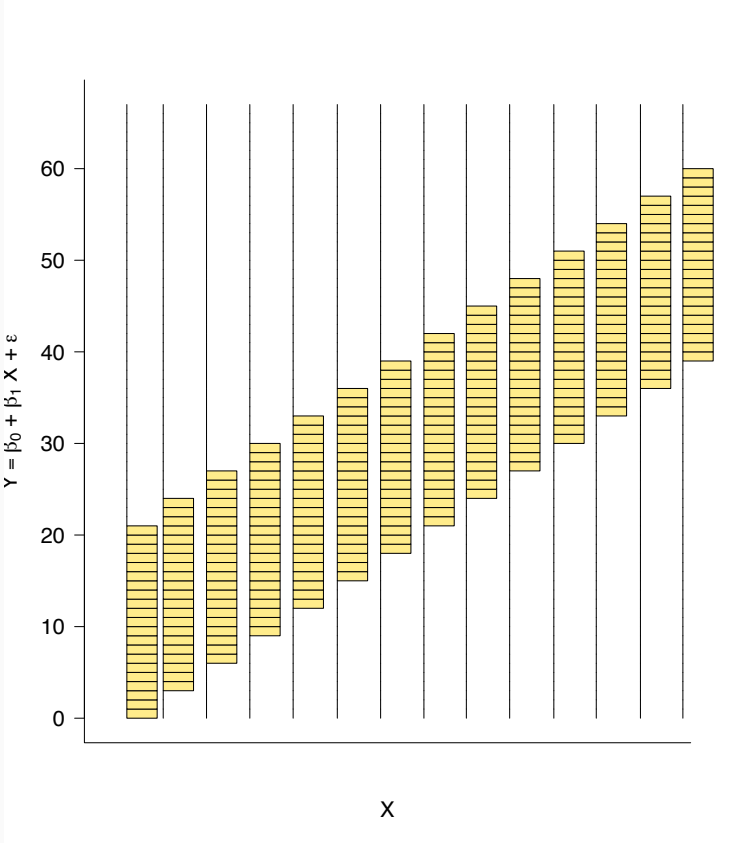


The case of heterogeneous errors



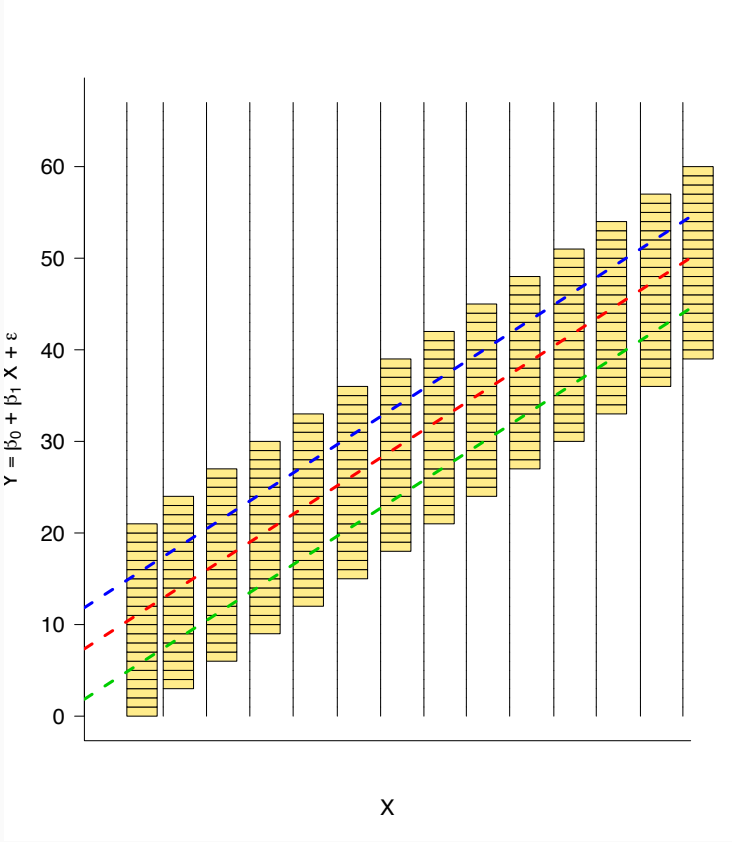
Location-scale model

The case of homogeneous (not normal) errors



Uniform error

The case of homogeneous (not normal) errors

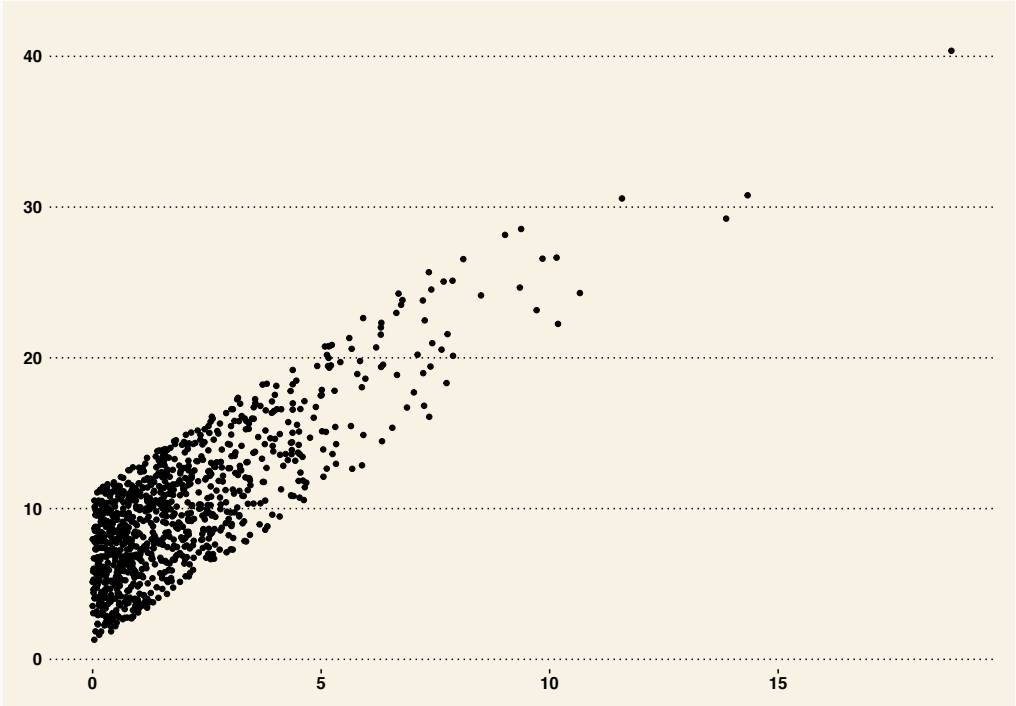


Location-shift model

The case of uniform errors

```r
# a uniform error model
set.seed(17)
n <- 1000
df_chi <- 2; beta0 <- 1; beta1 <- 2
x <- rchisq(n, df_chi)
error <- runif(n, 0, 10)
y <- beta0 + beta1 * x + error
# organize data into a dataframe
df <- data.frame(x, y)
# scatter plot
library(ggplot2); library(ggthemes)
g1 <- ggplot(data = df, aes(x, y)) +
  geom_point() + theme_wsj()
```


The case of uniform errors

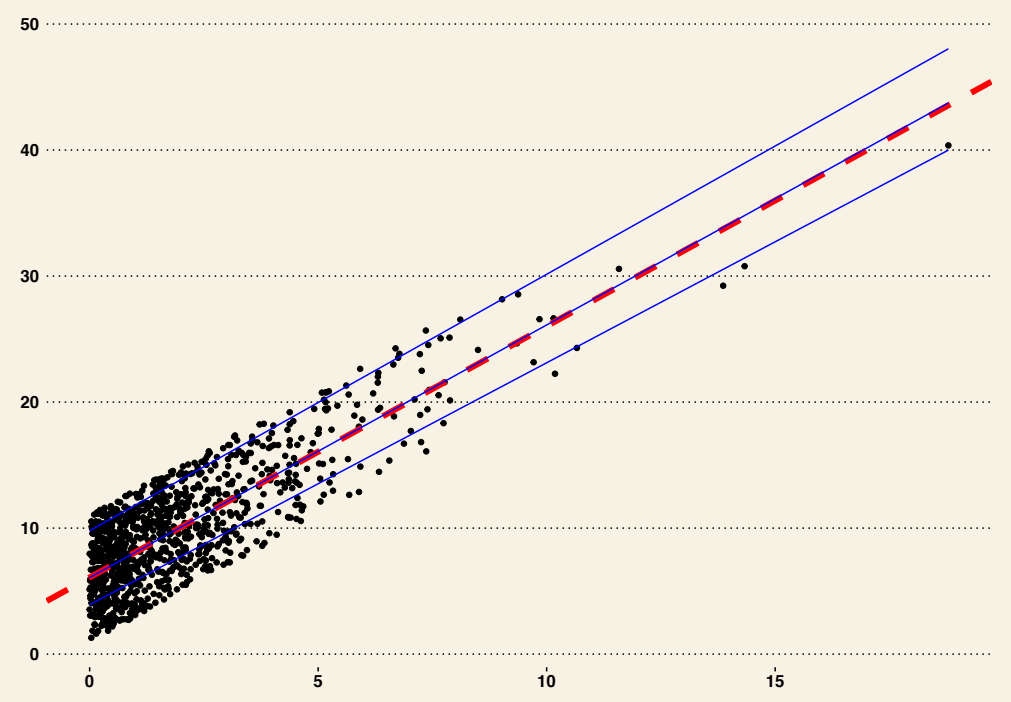


The case of uniform errors

```
lm_coefs <- coef(lm(y ~ x, data = df))

g2 <- g1 +
      geom_abline(intercept = lm_coefs[1],
                  slope = lm_coefs[2],
                  colour = "red",
                  linetype = "dashed", size = 2) +
      stat_quantile(quantiles = c(0.25, 0.5, 0.9),
                    colour = "blue")
```

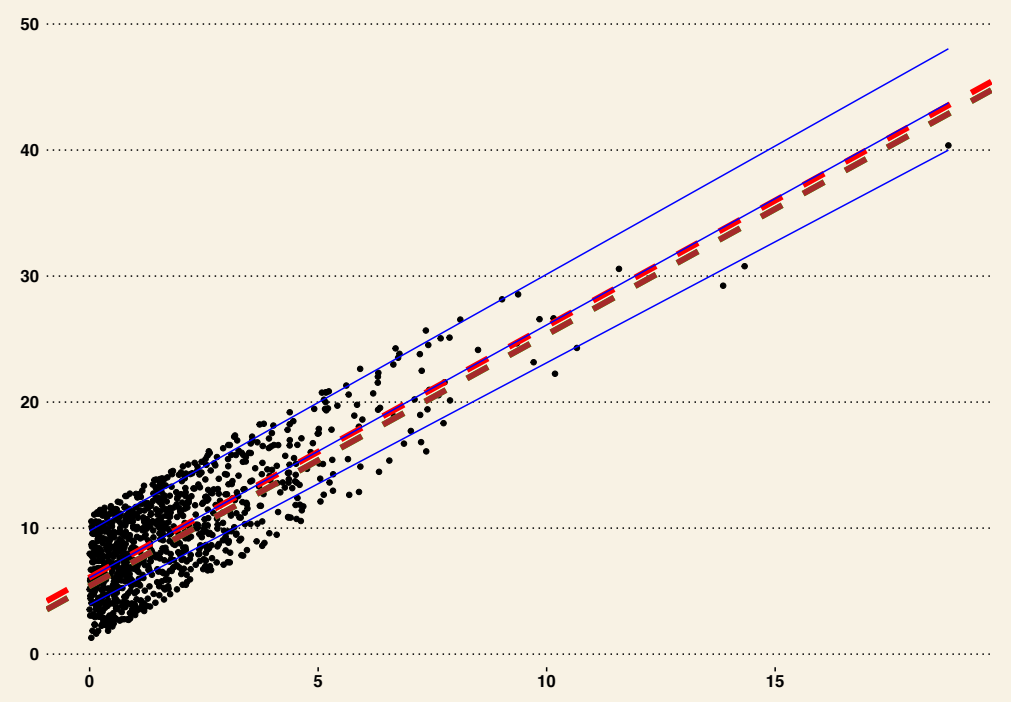

The case of uniform errors



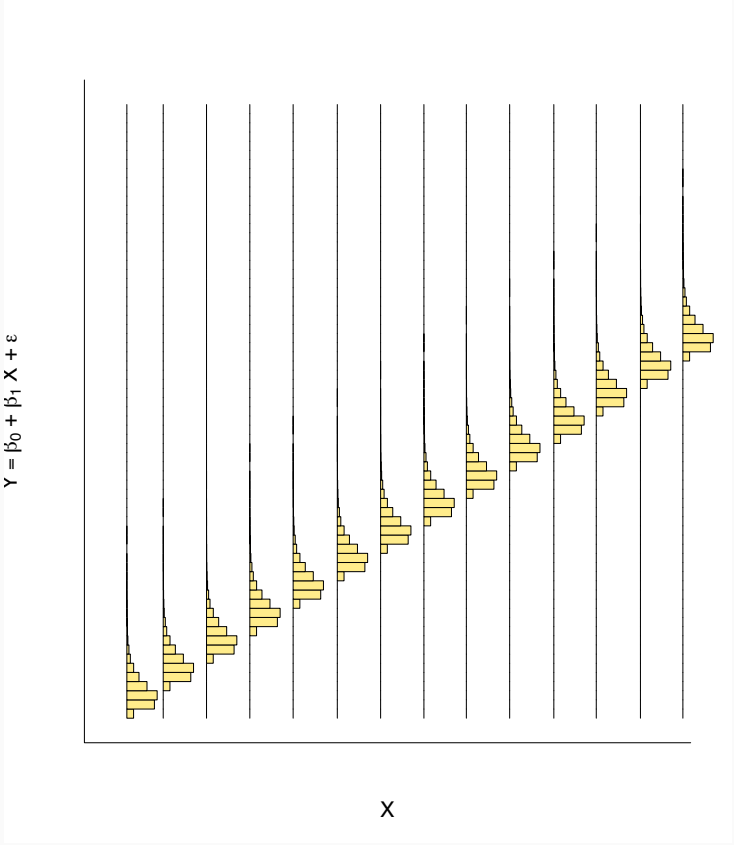
The case of uniform errors

```
g3 <- g2 +
  geom_abline(intercept = lm_coefs[1] +
                qnorm(0.25) *
                summary(lm(y ~ x, data = df))$sigma,
              slope = lm_coefs[2],
              colour = "green",
              linetype = "dashed", size = 2) +
  geom_abline(intercept = lm_coefs[1] +
                qnorm(0.9) *
                summary(lm(y ~ x, data = df))$sigma,
              slope = lm_coefs[2],
              colour = "brown",
              linetype = "dashed", size = 2)
```


The case of uniform errors

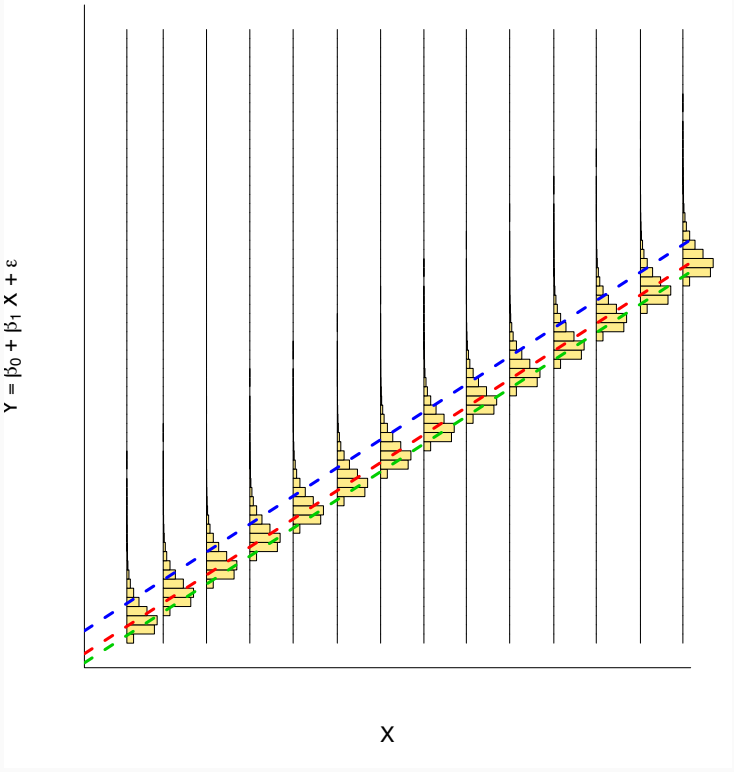


The case of homogeneous (not normal) errors



Lognormal error

The case of homogeneous (not normal) errors

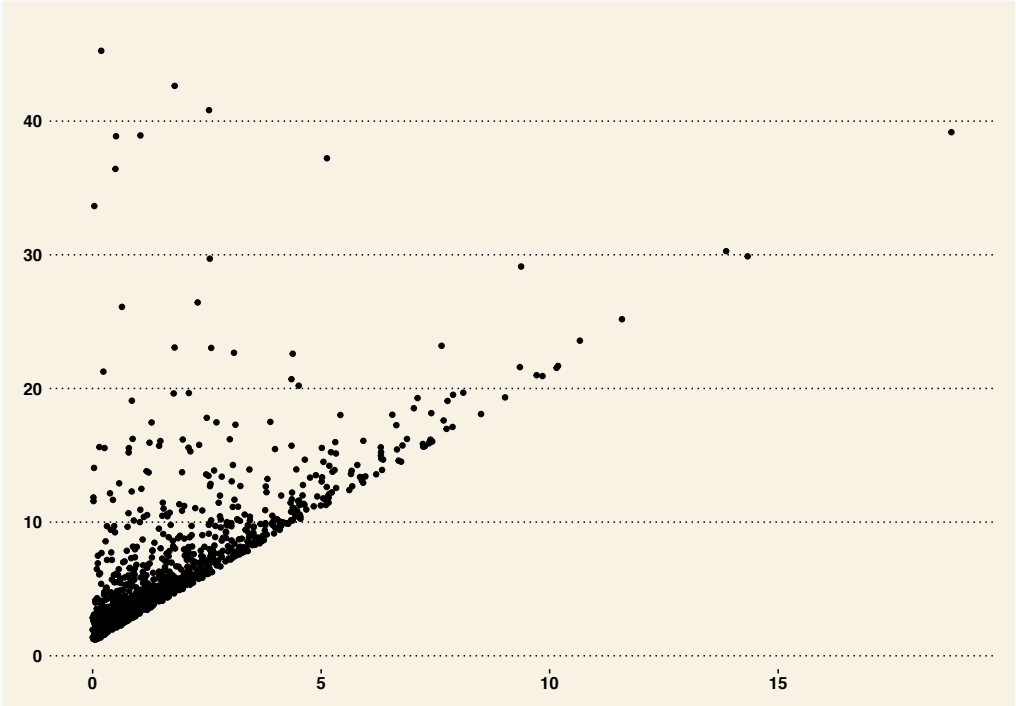


Location-shift model

The case of log-normal errors

```
# a lognormal error model
set.seed(17)
n <- 1000
df_chi <- 2; beta0 <- 1; beta1 <- 2
x <- rchisq(n, df_chi)
error <- rlnorm(n, 0, 1.25)
y <- beta0 + beta1 * x + error
# organize data into a dataframe
df <- data.frame(x, y)
# scatter plot
library(ggplot2); library(ggthemes)
g1 <- ggplot(data = df, aes(x, y)) +
  geom_point() + theme_wsj()
```


The case of log-normal errors

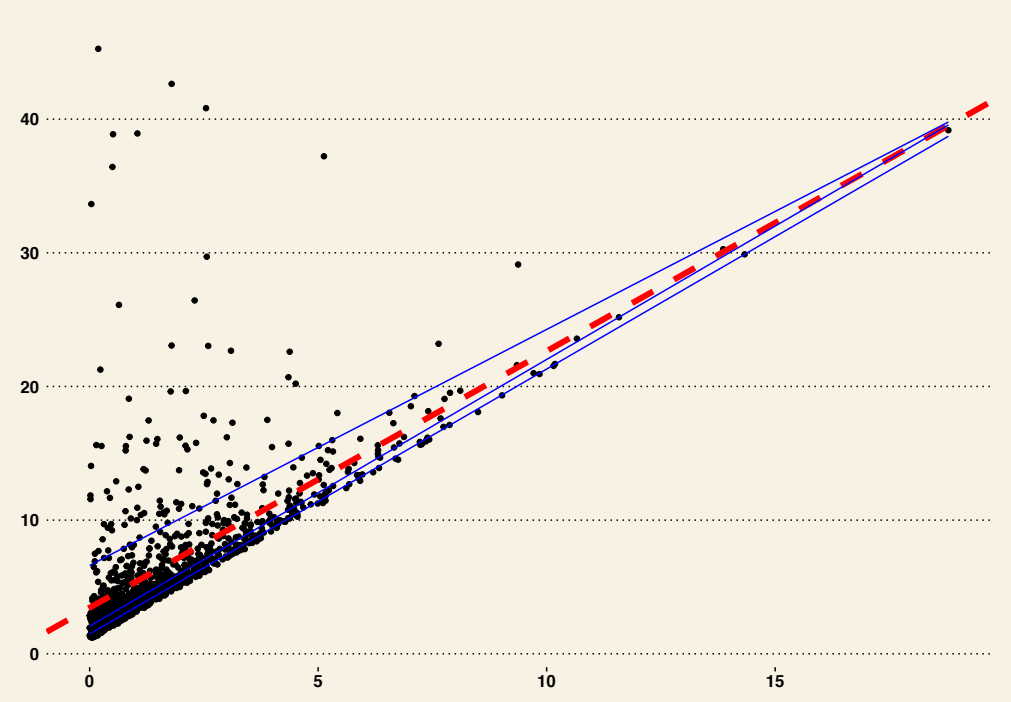


The case of log-normal errors

```
lm_coefs <- coef(lm(y ~ x, data = df))

g2 <- g1 +
      geom_abline(intercept = lm_coefs[1],
                  slope = lm_coefs[2],
                  colour = "red",
                  linetype = "dashed", size = 2) +
      stat_quantile(quantiles = c(0.25, 0.5, 0.9),
                    colour = "blue")
```

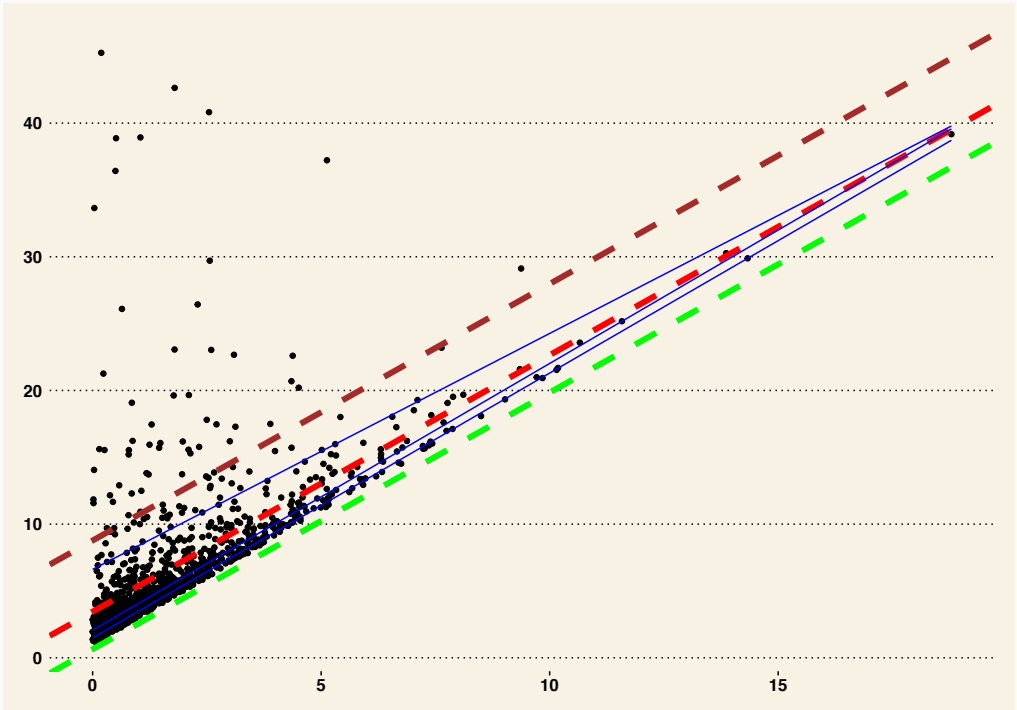

The case of log-normal errors



The case of log-normal errors

```
g3 <- g2 +
  geom_abline(intercept = lm_coefs[1] +
                qnorm(0.25) *
                summary(lm(y ~ x, data = df))$sigma,
              slope = lm_coefs[2],
              colour = "green",
              linetype = "dashed", size = 2) +
  geom_abline(intercept = lm_coefs[1] +
                qnorm(0.9) *
                summary(lm(y ~ x, data = df))$sigma,
              slope = lm_coefs[2],
              colour = "brown",
              linetype = "dashed", size = 2)
```


The case of log-normal errors

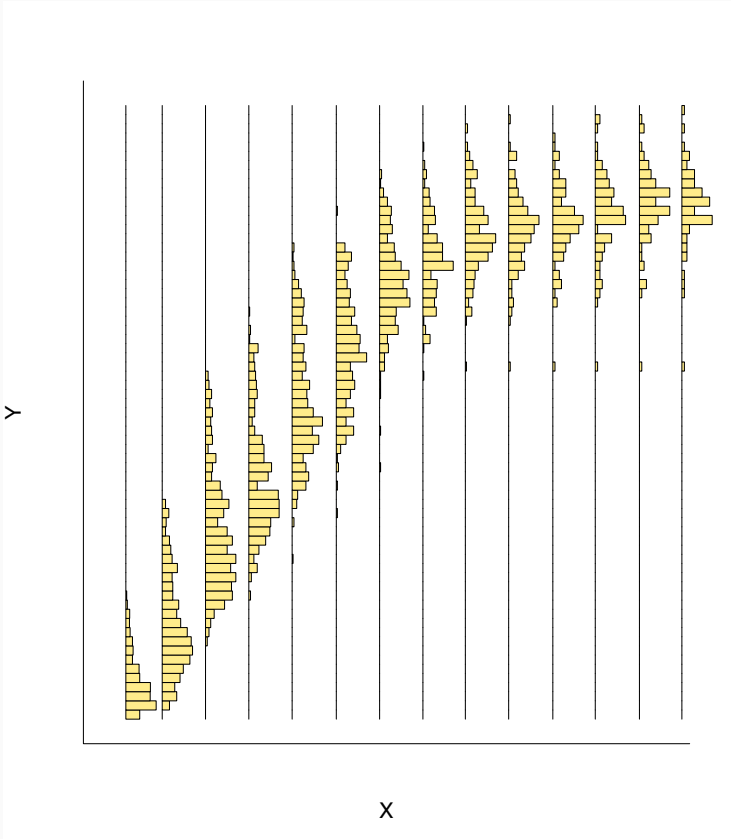


A more complex world

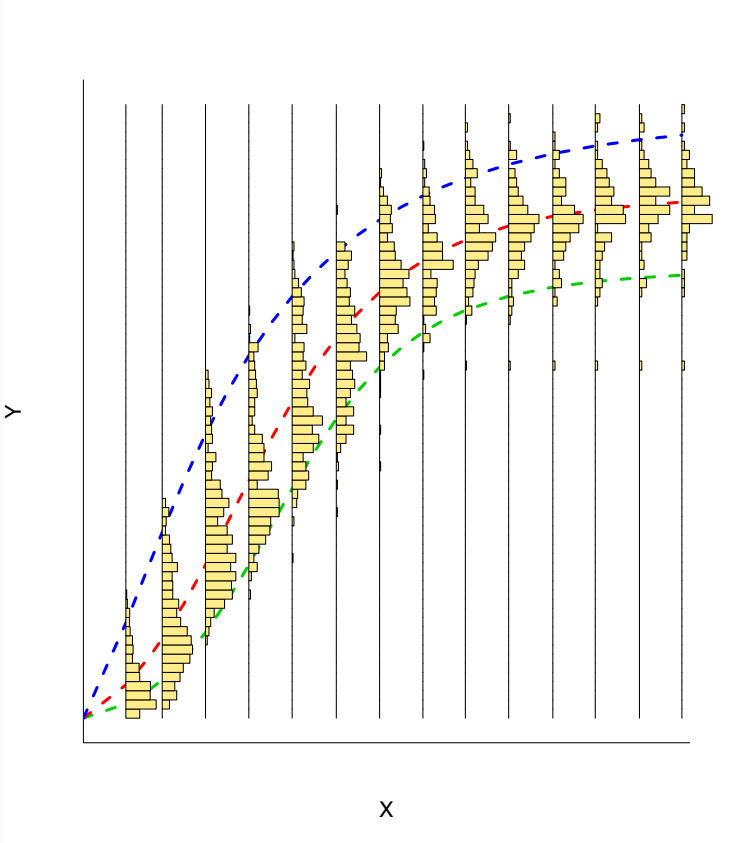
We live in a paradoxical world, where the only true safety, true though limited, comes from admitting both our uncertainty and the incompleteness with which we are able to meet it.

J. W. Tukey (1997)

A more complex world



A more complex world



Some possible approaches

- regression models for location, scale and shape
- quantile regression
- expectile regression

Regression models for location, scale and shape

- There are still assumptions on a specific error distribution but covariates exert effect not only on the mean
- A distribution for the response is specified, where (potentially) all parameters are related to predictors
- A simple example? Regression for mean and variance of a normal distribution where:

$$y_i = \eta_{i1} + exp(\eta_{i2}\epsilon_i)$$

with:

$$\epsilon_i \sim N(0, 1)$$

In such a model, we have:

- $E(y_i|\mathbf{z}_i) = \eta_{i1}$
- $Var(y_i|\mathbf{z}_i) = exp(\eta_{i2})^2$

Quantile regression

- no parametric assumptions for the error (and hence response) distribution
- estimation of separate models for different asymmetries $\tau \in [0, 1]$
- insted of $E(\epsilon_{i\tau} = 0)$, we have $P(\epsilon_{i\tau} \leq 0) = \tau$, i.e. the τ-quantile of the error term is 0
- the separate models are interpretable in terms of regression models for the quantiles of the response
- a dense set of quantiles completely characterizes the conditional distribution of the response

Expectile regression

- expectiles are a computationally alternative to quantiles
- how do you interpret an expectile?
- please, be patient until this afternoon

The (un)official history

Quantile regression timeline

A (not so short) history

1755

Jesuit Roger Joseph Boscovich

Problem of ellipticity of the earth

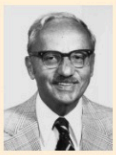
Being given a certain number of degrees, find the correction that must be made to each of them , supposing these three conditions are complied with: the first, that their differences shall be proportional to the differences between the versed sines of twice their latitudes; the second, that the sum of the positive corrections shall be equal to the sum of the negative ones; the third, that the sum of all the corrections, positive as well as negative, shall be the least possible, for the case where the first two conditions will be fulfilled.

1805

Adrien Marie Legendre

Nouvelles Méthods pour la détermination des orbites des comèetes

Appendix:
Sur la méthode des moindres quarrés



1947

George Bernard Dantzig

The simplex algorithm

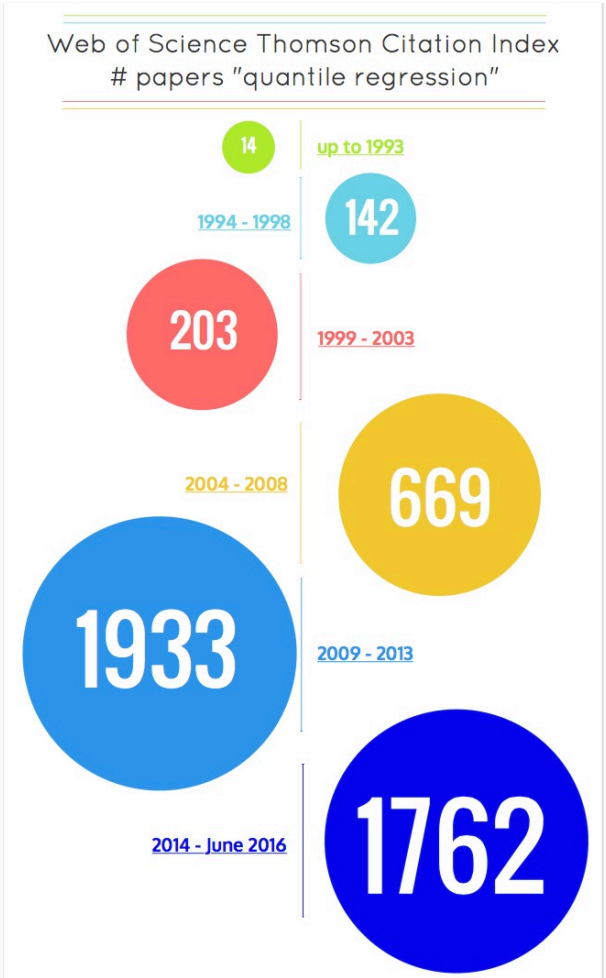
Linear programming and extension
Princeton University Press (1963)

Roger Koenker & Gig Basset

Regression quantiles
Econometrica

1978

Some (un)official statistics


Web of Science Thomson Citation Index
papers "quantile regression"
14
up to 1993
1994 - 1998
142
203
1999 - 2003
2004 - 2008
669
1933
2009 - 2013
2014 - June 2016
1762


Economic Applications of Quantile Regression
2002
2005
Quantile Regression
Roger Koenker
2007
QUANTILE REGRESSION
2012
Quantile Regression for Spatial Data
2013
Quantile Regression
Theory and Applications
WILEY
2016
2016
Investigations on Quantile Regression

Quantile regression software

- R:
  - quantreg package (Koenker)
  - 24 additional packages (ALDqr, bayesQR, BSquare, cdfquantreg, cmprskQR, cqrReg, expectreg, factorQR, GLDreg, hqreg, lqr, modQR, OutlierDC, plaqr, QICD, qrcm, qrjoint, qrLMM, qrNLMM, qrnn, quantregForest, quantregGrowth, quantreg.nonpar, rqPen)
- SAS: QUANTREG procedure
- Stata: qreg, sqreg, iqreg

and, recently, also:

- EViews7
- XL-Stat

Description of the 25 available R packages (1 of 3)

ALDqr	Quantile Regression Using Asymmetric Laplace Distribution
bayesQR	Bayesian quantile regression
BSquare	Bayesian Simultaneous Quantile Regression
cdfquantreg	Quantile Regression for Random Variables on the Unit Interval
cmprskQR	Analysis of Competing Risks Using Quantile Regressions
cqrReg	Quantile, Composite Quantile Regression and Regularized Versions
expectreg	Expectile and Quantile Regression
factorQR	Bayesian quantile regression factor models

Description of the 25 available R packages (2 of 3)

GLDreg	Fit GLD Regression Model and GLD Quantile Regression Model to Empirical Data
hqreg	Regularization Paths for Lasso or Elastic-Net Penalized Huber Loss Regression and Quantile Regression
lqr	Robust Linear Quantile Regression
modQR	Multiple-Output Directional Quantile Regression
OutlierDC	Outlier Detection using quantile regression for Censored Data
plaqr	Partially Linear Additive Quantile Regression
QICD	Estimate the Coefficients for Non-Convex Penalized Quantile Regression Model by using QICD Algorithm
qrcm	Quantile Regression Coefficients Modeling

Description of the 25 available R packages (3 of 3)

qrjoint	Joint Estimation in Linear Quantile Regression
qrLMM	Quantile Regression for Linear Mixed-Effects Models
qrNLMM	Quantile Regression for Nonlinear Mixed-Effects Models
qrnn	Quantile Regression Neural Network
quantreg	Quantile Regression
quantregForest	Quantile Regression Forests
quantregGrowth	Growth Charts via Regression Quantiles
quantreg.nonpar	Nonparametric Series Quantile Regression
rqPen	Penalized Quantile Regression

User guide

- A basic introduction - see external presentation

An useful property

Equivariance

Equivariance properties

Ability to use the same transformation rules when the data on the model are subject to a transformation

Transformation of variable scale is very common:

- to aid interpretation
- to attain a better model fit

Note: Buckinsky (1998) proposed to exploit the equivariance property to speed up the estimation process by reducing the number of simplex iterations

Equivariance

- scale equivariance
- shift or regression equivariance
- equivariance to repametrization of design
- **equivariance to monotone transformations**

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}$$

$$Q_\theta\left[\widehat{h(\mathbf{y})}|\mathbf{x}\right] = h\left[\hat{\beta}_0(\theta)\right] + h\left[\hat{\beta}_1(\theta)\right]\mathbf{x}$$

where $h(.)$ is a non decreasing function in $\Re$

Linear Equivariance

Linear equivariance

For any linear transformation of the response variable, both the conditional mean and the conditional quantiles can be exactly transformed

- conditional mean

$$E(a + b\mathbf{y}|\mathbf{x}) = a + bE(\mathbf{y}|\mathbf{x})$$

- conditional quantiles
  - if $b > 0$:

$$Q_\theta(a + b\mathbf{y}|\mathbf{x}) = a + bQ_\theta(\mathbf{y}|\mathbf{x})$$

  - if $b < 0$:

$$Q_\theta(a + b\mathbf{y}|\mathbf{x}) = a + bQ_{1-\theta}(\mathbf{y}|\mathbf{x})$$

Monotone transformations (1/2)

- log–transformation is a typical nonlinear transformation, used:
  - to address the right–skewness of a distribution
  - to model a covariate's effect in relative terms (e.g. percentage change)
- it is not possible to obtain the conditional mean of the response in absolute terms starting from the conditional mean on the log–scale

$$E\left[log(\mathbf{y})|\mathbf{x}\right] \neq log(E\left[\mathbf{y}|\mathbf{x}\right])$$

$$E\left[y_i|x_i\right] \neq e^{E[log(\mathbf{y})|\mathbf{x}]}$$

Note: It would be a mistake to use the $log(y)$ results to make conclusions about the distribution of Y (though this is a widely used practice)

Monotone transformations (2/2)

Monotone transformation

Transformation that preverses order

- given $y < y'$ then $h(y) < h(y')$

In general:

$$E\left[h(\mathbf{y})|\mathbf{x}\right] \neq h(E\left[\mathbf{y}|\mathbf{x}\right])$$

while:

$$Q_\theta\left[h(\mathbf{y})|\mathbf{x}\right] = h(Q_\theta\left[\mathbf{y}|\mathbf{x}\right])$$

Note: this property follows immediately from the monotone equivariance property of univariate quantiles

A note on inference

Main approaches to inference in QR

- Small sample theory
  (Koenker and Basset, 1978)
  "*The practical of this theory would entail a host of hazardous assumptions and an exhausting computational effort*" (Koenker, 2005)
- Asymptotic theory
  (Koenker and Basset, 1978, 1982a,b)
- Rank–based theory
  (Gutenbrunner and Jureckova, 1992) (Gutenbrunner , 1993)
- Resampling methods
  (Parzen , 1994) (He and Hu, 2002) (Kocherginsky, 2003; Kocherginsky , 2005)

Main approaches to inference in QR

- Small sample theory
  (Koenker and Basset, 1978)
  "*The practical of this theory would entail a host of hazardous assumptions and an exhausting computational effort*" (Koenker, 2005)
- Asymptotic theory
  (Koenker and Basset, 1978, 1982a,b)
- Rank–based theory
  (Gutenbrunner and Jureckova, 1992) (Gutenbrunner, 1993)
- Resampling methods
  (Parzen, 1994) (He and Hu, 2002) (Kocherginsky, 2003; Kocherginsky, 2005)

Asymptotic theory

$$Q_\theta(\hat{y}|x) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)x$$

"under mild regularity conditions"

$$\Downarrow$$

Asymptotic distribution of QR estimator:

$$\frac{\hat{\beta}(\theta) - \beta(\theta)}{SE(\hat{\beta}(\theta))} \rightarrow N(0,1)$$

$SE(\hat{\beta}(\theta))$ depends on the error distribution (i.e. the error distribution affects the variance–covariance matrix of the QR estimator)

- standard errors are simpler and easier to describe under the i.i.d. model
- it is quite complex to deal with the ni.i.d. case, as the errors no longer have a common distribution

The bootstrap procedure

The bootstrap procedure is usally preferable to the asymptotic for two reasons:

- when the assumptions for the asymptotic procedure do not hold
- even if the required assumptions are satisfied, the solutions for the standard errors of the construscted and skewness shifts, are complicated to compute

The boostrap procedure offers the flexibility to obtain the standard error and confidence interval for any estimates and combinations of estimates

A note on robustness

A note on robustness

- QR estimates are not sensitive to outliers in *Y*: if we modify the value of the response variable for a data point lying above (or below) the fitted QR lines, without changing the sign of the corresponding sign
- QR estimator can be very sensitive to outliers in the explanatory variables (He et al, 1990)
- Several proposals in the literature attempt to attain a more robust for of QR: (Rousseauw and Hubert, 1999), Giloni et al. (2006), and more recently Neykov (2012)

Estimation: technical details

Estimation: technical details

Conditional mean and conditional quantiles

On optimal criteria

- QR extends regression analysis to the study of the whole conditional distribution of the response
- QR is to classical regression what quantiles are to mean in terms of describing locations of a distribution

Unconditioanl mean and median (nothing of new)
Let Y a generic random variable:

- Mean (and its objective function): $\mu = \text{argmin}_c \, E(Y - c)^2$
- Median (and its objective function): $Me = \text{argmin}_c \, E|Y - c|$

Note: with $\hat{\mu}$ and $\hat{Me}$ we denote the two sample estimators

On optimal criteria

Quantiles as particular locations of the distribution

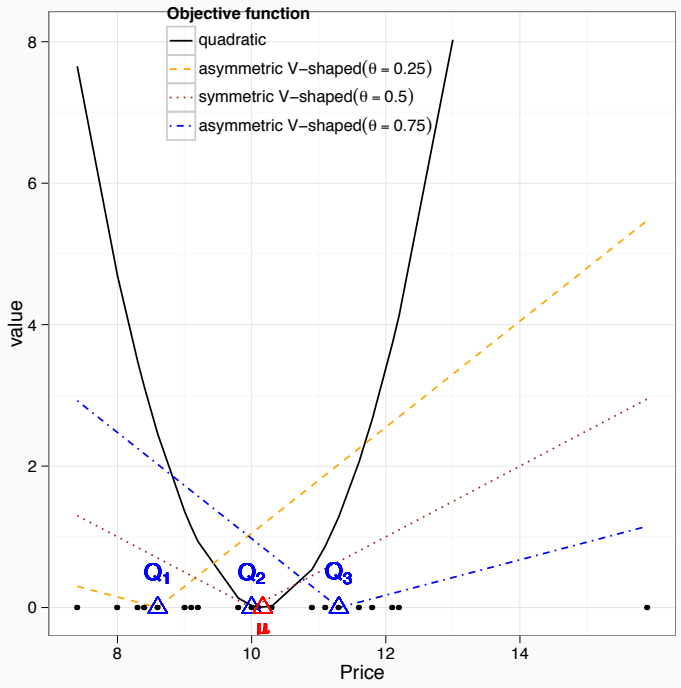
$$q_\theta = \underset{c}{\text{argmin}} \, E[\rho_\theta(Y - c)]$$

where $\rho_\theta(.)$ denotes the following loss function:

$$\begin{aligned} \rho_\theta(y) &= [\theta - I(y < 0)]y \\ &= [(1 - \theta)I(y \leq 0) + \theta I(y > 0)]|y| \end{aligned}$$

It is an asymmetric absolute loss function: a weighted sum of absolute deviations, where a $(1 - \theta)$ weight is assigned to the negative deviations and a weight θ is used for the positive deviations

On optimal criteria



On optimal criteria

- In case of a discrete random variable Y with probability distribution $f(y) = P(Y = y)$, we have:

$$q_\theta = \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y - c)]$$

$$= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \sum_{y \le c} |y - c| f(y) + \theta \sum_{y > c} |y - c| f(y) \right\}$$

- In case of a continuous random variable Y with probability density function $f(y) = P(Y = y)$, we have:

$$q_\theta = \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y - c)]$$

$$= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \int_{-\infty}^{c} |y - c| f(y) d(y) + \theta \int_{c}^{+\infty} |y - c| f(y) d(y) \right\}$$

Note 1: $\hat{q}_\theta$, for $\theta \in [0, 1]$, denotes the sample estimator for the conditional quantile

Note 2: for $\theta = 0.5$ we obtain the median solution

Solution for $q_{\theta=0.5}$: the case of the median

Assuming, without loss of generality, that Y is a continuous random variable, the expected value of the absolute sum of deviations can be formulated as:

$$\begin{aligned} E|Y-c| &= \int_{y\in\mathcal{R}} |y-c| f(y)dx \\ &= \int_{y<c} |y-c| f(y)dy + \int_{y>c} |y-c| f(y)dy \\ &= \int_{y<c} (c-y) f(y)dy + \int_{y>c} (y-c) f(y)dy \end{aligned}$$

Since the absolute value is a convex function, differentiating $E|Y-c|$ with respect to c and setting the partial derivatives to zero will lead the solution for the minimum:

$$\frac{\partial}{\partial c} E|Y-c| = 0$$

Solution for $q_{\theta=0.5}$: the case of the median

Then, applying the derivative and integrating per part:

$$\left\{ (c-y)f(y)\Big|_{-\infty}^{c} + \int_{y<c} \frac{\partial}{\partial c}(c-y)f(y)dy \right\} +$$
$$\left\{ (y-c)f(y)\Big|_{c}^{+\infty} + \int_{y>c} \frac{\partial}{\partial c}(y-c)f(y)dy \right\} = 0$$

Taking into account that $f(-\infty) = f(+\infty) = 0$ for a well-defined probability density function, the integrand restricts in $y = c$:

$$\left\{ \underbrace{(c-y)f(y)\Big|_{y=c}}_{\text{= 0 when y = c}} + \int_{y<c} f(y)dy \right\} + \left\{ \underbrace{(y-c)f(y)\Big|_{y=c}}_{\text{= 0 when y = c}} - \int_{y>c} f(y)dy \right\}$$

Note: Our interest is in $y = c$ (E|Y - c | is minimized)

Solution for $q_{\theta=0.5}$: the case of the median

Therefore, we have:

$$\int_{y<c} f(y)dy + \int_{y>c} f(y)dy = 0$$

namely:

$$F(c) - [1 - F(c)] = 0$$

and thus:

$$2F(c) - 1 = 0 \implies F(c) = \frac{1}{2} \implies c = Me$$

Solution for the generic q_θ

The solution does not change by multiplying the two components of $E|Y - c|$ by a costant θ and $(1 - \theta)$, respectively:

$$\frac{\partial}{\partial c} E[\rho_\theta(Y - c)] = \frac{\partial}{\partial c} \left\{ (1-\theta) \int_{-\infty}^{c} |y - c| f(y)dy + \theta \int_{c}^{+\infty} |y - c| f(y)dy \right\}$$

Repeating the above argument, we easily obtain:

$$\frac{\partial}{\partial c} E[\rho_\theta(Y - c)] = (1-\theta)F(c) - \theta(1 - F(c)) = 0$$

and then q_θ as the solution of the minimization problem:

$$F(c) - \theta F(c) - \theta + \theta F(c) = 0 \implies F(c) = \theta \implies c = q_\theta$$

Technical details

Technical details

Conditional mean and conditional quantiles

Conditional mean and conditional quantiles

- by replacing the sorting with optimization, the above line of reasoning generalizes easily to the regression setting.
- denoting with Y the response variable and with $\mathbf{X}$ the set of predictor variables

Estimation of the conditional mean function $\mu(\mathbf{x}_i, \beta) = E[Y|\mathbf{X} = \mathbf{x}_i]$

$$\hat{\mu}(\mathbf{x}_i, \beta) = \underset{\mu}{\operatorname{argmin}}\, E[Y - \mu(\mathbf{x}_i, \beta)]^2$$

Estimation of the conditional quantile function

$$\hat{q}_Y(\theta, \mathbf{X}) = \underset{Q_Y(\theta,\mathbf{X})}{\operatorname{argmin}}\, E[\rho_\theta(Y - Q_Y(\theta, \mathbf{X}))]$$

Conditional mean and conditional quantiles: the linear case

- When $\mu(\mathbf{x}_i, \beta) = \mathbf{x}_i^\top \beta$, we have:

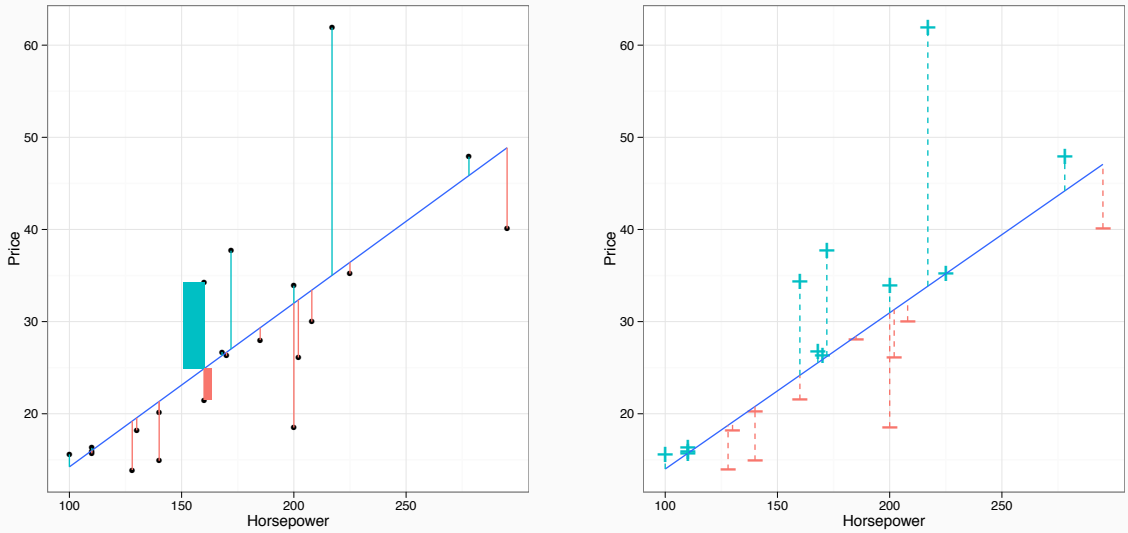
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, E[Y - \mathbf{x}_i^\top \beta)]^2$$

- Similarly, in case of linear quantile functions:

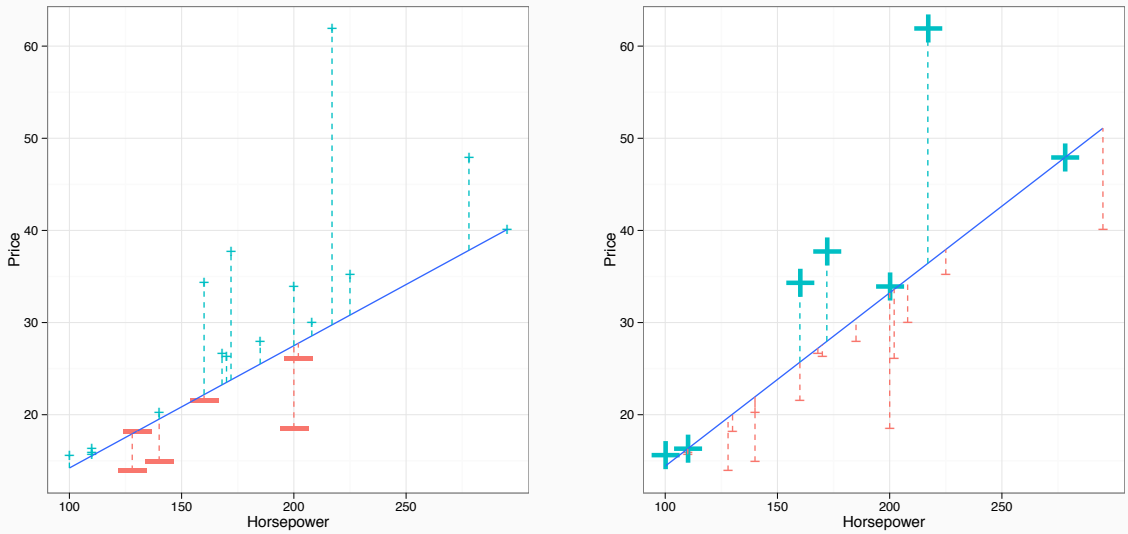
$$\hat{\beta}(\theta) = \underset{\beta}{\operatorname{argmin}}\, E[\rho_\theta(Y - \mathbf{X}\beta)]$$

Note: the (θ)-notation denotes that the parameters and the corresponding estimators are for a specific quantile θ

Geometrical interpretation: mean vs median



Geometrical interpretation: $q_{0.25}$ vs $q_{0.75}$



Technical details

Technical details

The QR problem through linear programming

A very short introduction to linear programming

- Linear programming is an extremely flexible tool widely used in different application fields
- Linear programming is a subset of mathematical programming facing the efficient allocation of limited resources to known activities
- The allocation aims to minimize a cost or maximize a profit

A very short introduction to linear programming

- The variables:

$$x_i \geq 0 \qquad i = 1, \ldots, n,$$

  whose values are to be decided are referred to as *decisional variables*
- The aim of a linear programming problem is to find a vector $\mathbf{x}^* \in \mathbb{R}^n_+$ minimizing (or maximizing) the value of a given linear function
- The vector $\mathbf{x}^*$ is detected among all vectors $\mathbf{x} \in \mathbb{R}^n_+$ that satisfy a given system of linear equations and inequalities
- The role of linearity is twofold: 1) the objective function (the quality of the plan) is measured through a linear function of the considered quantities; 2) feasible plans are restricted by linear constraints (inequalities)
- The linearity of some models is determined by the typical properties of the problem
- Some nonlinear problems (as in the case of QR) can be linearized by a proper use of mathematical transformations
- The representation of a problem in terms of linear programming ensures that efficient procedures are available for computing the solutions

Linear programming requirements (1/3)

- The n decision variables are non-negative:

$$x_i \geq 0 \quad i = 1, \ldots, n.$$

Note:

In case of variables unrestricted in sign, a simple trick can be exploited:

$$\begin{aligned} x &= [x]^+ - [-x]^+ \\ [x]^+ &\geq 0 \\ [-x]^+ &\geq 0, \end{aligned}$$

where $[x]^+$ denotes the non-negative part of x

We have:

- if $[x]^+ > 0$, then $[-x]^+ = 0$ and $x = [x]^+ > 0$
- for $[-x]^+ > 0$, then $[x]^+ = 0$ and $x = -[-x]^+ < 0$
- if $[x]^+ = [-x]^+ = 0$, then $x = 0$

Linear programming requirements (2/3)

- the criterion for choosing the optimal values of the decision variables (the objective function) is a linear function of the same variables:

$$z = \sum_{i=1}^{n} c_i x_i = \mathbf{cx}.$$

Note:

The conversion from a minimization to a maximization problem is trivial: maximize z is equivalent to minimize $-z$

Linear programming requirements (3/3)

- the m constraints regulating the process can be expressed as linear equations and/or linear inequalities written in terms of the decision variables

$$a_1x_1 + \ldots + a_ix_i + \ldots + a_nx_n \left\{ \begin{array}{c} \leq \\ = \\ \geq \end{array} \right\} b.$$

Some technical tricks for expressing constraints (1/3)

Note:

It is easy to convert constraints from one form to another

- For example, an inequality constraint:

$$a_1x_1 + \ldots + a_ix_i + \ldots + a_nx_n \leq b.$$

  can be converted to a greater than or equal constraints simply by multiplying it by -1

Some technical tricks for expressing constraints (1/3)

- An inequality constraint can be converted to an equality constraint by adding a non-negative variable (*slack* variable):

$$a_1x_1 + \ldots + a_ix_i + \ldots + a_nx_n + w = b, \quad w \geq 0.$$

Some technical tricks for expressing constraints (1/3)

- An equality constraint:

$$a_1x_1 + \ldots + a_ix_i + \ldots + a_nx_n = b$$

can be converted to an inequality form through the introduction of two inequality constraints:

$$\begin{aligned} a_1x_1 + \ldots + a_ix_i + \ldots + a_nx_n &\leq b \\ a_1x_1 + \ldots + a_ix_i + \ldots + a_nx_n &\geq b. \end{aligned}$$

A generic linear programming problem

Problem with n decision variables and m constraints

Standard form:

$$\begin{aligned} \text{minimize} \quad & \mathbf{cx} \\ \text{subject to} \quad & \mathbf{Ax} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

- $\mathbf{c}_{[n]}$ contains the costs for the n decision variables
- $\mathbf{A}_{[m \times n]}$ and $\mathbf{b}_{[m]}$ contain the coefficients corresponding to the m constraints

Note:

- The standard form poses the inequalities as a less than or equal form and requires the non-negativity for the decision variables
- The equational form (useful for the simplex algorithm) is obtained through the introduction of "slack" variables.
- In this last form, the vector $\mathbf{x}$ contains both the decision variables and the slack variables

Geometric interpretation

- A linear equation corresponds to a hyperplane
- An inequality divides the n–dimensional space into two half–spaces, one in which the inequality is satisfied and the other in which it is not
- The constraints $\mathbf{x} \geq \mathbf{0}$ restricts $\mathbf{x}$ to $\mathbb{R}^n_+$, that is the positive quadrant in n–dimensional space: in $\mathbb{R}^2$ it is a quarter of the plane, in $\mathbb{R}^3$ it is an eighth of the space, and so on
- The constraints $\mathbf{Ax} \leq \mathbf{b}$ produce m additional halp–spaces
- The feasible set consists of the intersection of the above mentioned $m + n$ half–spaces: n corresponding to the decision variables and m to the constraints
- The cost function $\mathbf{cx}$ produces a family of parallel planes: the place $\mathbf{cx} = \textit{constant}$ corresponds to the plane whose cost is equal to *constant*
- When *constant* varies, the plane sweeps out the n–dimensional space
- The optimal solution $\mathbf{x}^*$ is the $\mathbb{R}^n_+$ point, that is the n–dimensional vector, that ensures the lowest cost lying in the feasible set

Geometric interpretation

Note:

- The feasible set can be empty, unbounded or bounded
- A solution x is feasible if it satisfies all the constraints
- The solution x is optimal (x^*) when the objective function is minimal
- The optimal vector $\mathbf{x}^*$ is the feasible set whose associated cost is minimal
- A problem that has no feasible solution is called infeasible
- A problem with arbitrarily larger objective values is unbounded

Dual formulation

- Associated with every linear program is its dual formulation
- *"Duality theory is the most important theoretical result about linear programs"* (Matousek e Gardner, 2007)
- The dual problem formulation uses the same $\mathbf{A}$ e $\mathbf{b}$ but reverses the point of view: the cost vector $\mathbf{c}$ and the constraint vector $\mathbf{b}$ are indeed switched to the dual profit vector $\mathbf{b}$ and to the dual constraint vector $\mathbf{b}$
- The dual unknown (decision variables), $\mathbf{y}$, is now a vector with m components
- The n constraints are represented by $\mathbf{yA} \geq \mathbf{c}$

Dual formulation

Dual formulation for the linear programming problem

$$\begin{aligned} \text{maximize} \quad & \mathbf{by} \\ \text{subject to} \quad & \mathbf{yA} \geq \mathbf{c} \\ & \mathbf{y} \geq \mathbf{0}. \end{aligned}$$

- Linear programs come then in primal/dual pairs
- Each feasible solution for one of these two linear programs gives a bound on the optimal objective value for the other
- The dual formulation is sometime easier to solve

Primal problem and dual problem

- *weak duality theorem*: the dual problem provides upper bounds for the primal problem

$$\mathbf{cx} \leq \mathbf{by};$$

- *strong duality theorem*: if the primal problem has an optimal solution, then the dual also has an optimal solution

$$\mathbf{cx}^* = \mathbf{by}^*$$

Methods for solving the LP problem

- The simplex algorithm (Dantzig, 1947) is the most famous method for solving a LP problem: it is an iterative algorithm, starting from a solution that satisfies the constraints and the non-negativities posed by the decision variables
- It looks for a new and better solution, that is a solution characterized by a lower (primal) or higher (dual) objective function value: the process iterates until a solution that cannot be further improved is reached
- The planes corresponding to the cost function to be minimized in the primal formulation (the profit function in the dual formulation) move up (down) until they intersect the feasible set
- The first contact must occur along the boundary
- The simplex algorithm essentially consists of movements along the edges of the feasible set: starting from an initial solution, the procedure goes from corner to corner of the feasible set until it finds the corner with the lowest (highest) associate cost (profit)

Methods for solving the LP problem

- Koenker e D'Orey (1987) introduced a variant of the efficient version of the simplex algorithm, proposed by Barrodale e Roberts (1974), to compute conditional quantiles
- The simplex algorithm is the default option in most of the QR software
- A completely different method approaches the solution from the interior of the feasible set rather than on its boundary (Karmakar, 1984)
- Interior–point methods have been shown to be competitive in case of very large problems
- Portnoy e Koenker (1997) proposed the use of interior–point methods for QR showing their efficiency in case of datasets with a large number of units
- The heuristic approach known as *finite smoothing algorithm (Chen 2004, 2007) is faster and more accurate for approximating the original problem with respect to interior–point method in presence of a large number of covariates*

The LP formulation of the QR problem

- L_1 regression, median regression, is a natural extension of the sample median when the response is conditioned on the covariates
- The LP formulation for the conditional median is shown first for the simple regression model and then for the multiple regression model

The LP formulation of the QR problem

L_1 criterion for the two–variables problem

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |\beta_0 + \beta_1 x_i - y_i|.$$

Although this cost function is not linear, a simple trick allows us to make it linear, at the price of introducing extra variables:

$$\begin{aligned} &\text{minimize} && \textstyle\sum_{i=1}^{n} e_i && \\ &\text{subject to} && e_i \geq \beta_0 + \beta_1 x_i - y_i && i = 1, \ldots, n \\ & && e_i \geq -(\beta_0 + \beta_1 x_i - y_i) && i = k+1, \ldots, n. \end{aligned}$$

Each e_i is an auxiliary variable standing for the error at the i–th point. The constraints guarantees that:

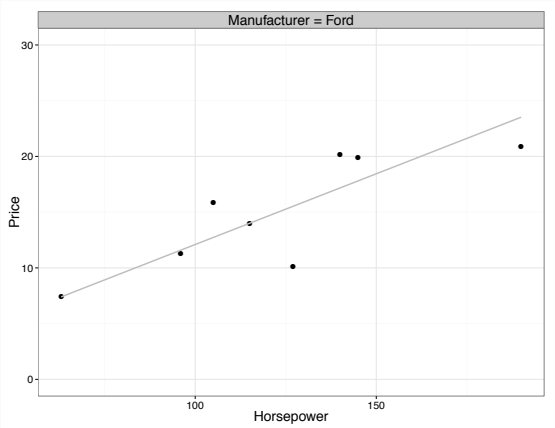
$$e_i \geq \max\{\beta_0 + \beta_1 x_i - y_i, -(\beta_0 + \beta_1 x_i - y_i)\} = |\beta_0 + \beta_1 x_i - y_i|.$$

- In an optimal solution, each of these inequalities has to be satisfied with equality; otherwise, we could decrease the corresponding e_i: the optimal solution thus yields a line minimizing the initial problem (without the additional variables)

- Hence, to solve the L_1 problem, it suffices to solve the equivalent LP problem

The LP formulation of the QR problem

Example: 8 Ford cars (dataset Cars93)



Horsepower	63	127	96	105	115	145	140	190
Price	7.4	10.1	11.3	15.9	14.0	19.9	20.2	20.9

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 = -0.60 \\ \hat{\beta}_1 = 0.13 \end{bmatrix}$$

The LP formulation of the QR problem

The L_1 regression problem is solved by finding the optimal solution to the following LP problem:

$$
\begin{array}{lrrrrrrrrrl}
\text{minimize} & \sum_{i=1}^{10} e_i & & & & & & & & & \\
\text{subject to} & - & \beta_0 - 63\,\beta_1 & -e_1 & & & & & & & \leq -7.4 \\
& - & \beta_0 - 127\,\beta_1 & & -e_2 & & & & & & \leq -10.1 \\
& - & \beta_0 - 96\,\beta_1 & & & -e_3 & & & & & \leq -11.3 \\
& - & \beta_0 - 105\,\beta_1 & & & & -e_4 & & & & \leq -15.9 \\
& - & \beta_0 - 115\,\beta_1 & & & & & -e_5 & & & \leq -14.0 \\
& - & \beta_0 - 145\,\beta_1 & & & & & & -e_6 & & \leq -19.90 \\
& - & \beta_0 - 140\,\beta_1 & & & & & & & -e_7 & \leq -20.2 \\
& - & \beta_0 - 190\,\beta_1 & & & & & & & -e_8 & \leq -20.9
\end{array}
$$

. . . follows on next page . . .

The LP formulation of the QR problem

. . . follows from previous page:

$$
\begin{array}{lrrrrrrrrrl}
\text{minimize} & \sum_{i=1}^{10} e_i & & & & & & & & & \\
\text{subject to} & \beta_0 + 63\,\beta_1 & -e_1 & & & & & & & & \leq 7.4 \\
& \beta_0 + 127\,\beta_1 & & -e_2 & & & & & & & \leq 10.1 \\
& \beta_0 + 96\,\beta_1 & & & -e_3 & & & & & & \leq 11.3 \\
& \beta_0 + 105\,\beta_1 & & & & -e_4 & & & & & \leq 15.9 \\
& \beta_0 + 115\,\beta_1 & & & & & -e_5 & & & & \leq 14.0 \\
& \beta_0 + 145\,\beta_1 & & & & & & -e_6 & & & \leq 19.9 \\
& \beta_0 + 140\,\beta_1 & & & & & & & -e_7 & & \leq 20.2 \\
& \beta_0 + 190\,\beta_1 & & & & & & & & -e_8 & \leq 20.9 \\
& & e_1, & e_2, & e_3, & e_4, & e_5, & e_6, & e_7, & e_8 & \geq 0.
\end{array}
$$

The LP formulation of the QR problem

L_1 criterion for the p–variables problem

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

where:

- $\mathbf{y}_{[n]}$ is the vector of responses
- $\mathbf{X}_{[n \times p]}$ is the regressor matrix
- $\boldsymbol{\beta}_{[p]}(\theta)$ is the vector of unknown parameters for the generic conditional quantile θ

Note: In the following, the simpler notation $\boldsymbol{\beta}$ is used to refer to the conditional median case ($\theta = 0.5$)

The LP formulation of the QR problem

Let us denote by $[x]_+$ the non negative part of x. By posing:

$$s_1 = [y - X\beta]_+$$
$$s_2 = [X\beta - y]_+$$

the original L_1 problem can be formulated as:

$$\min_{\beta}\{1^\top s_1 + 1^\top s_2 | y = X\beta + s_1 - s_2, \{s_1, s_2\} \in \mathbb{R}^n_+\}.$$

Furthermore, let:

$$B = [X - XI - I],$$

and:

$$\psi = \begin{bmatrix} [\beta]_+ \\ [-\beta]_+ \\ [y - X\beta]_+ \\ [X\beta - y]_+ \end{bmatrix}$$

$$d = \begin{bmatrix} 0_{[p]} \\ 0_{[p]} \\ 1_{[n]} \\ 1_{[n]} \end{bmatrix}$$

The LP formulation of the QR problem

Such reformulation of the problem leads to a standard linear programming problem

Primal formulation (equational form)

$$\begin{aligned} \underset{\psi}{\text{minimize}} \quad & d^\top \psi \\ \text{subject to} \quad & B\psi = y \\ & \theta \geq 0. \end{aligned}$$

Dual formulation (equational form)

$$\begin{aligned} \underset{d}{\text{maximize}} \quad & y^\top z \\ \text{subject to} \quad & B^\top z \leq d. \end{aligned}$$

Note: theorem ensuring that the solutions of such a minimization problem have to be searched in the corners of the simplex

The LP formulation of the QR problem

The above problem can be reformulated as follows:

$$\max_{z}\{y^\top z | X^\top z = 0, z \in [-1,+1]^n\}$$

In fact, the equality:

$$X^\top z = 0$$

can be transformed as follows:

$$\tfrac{1}{2}X^\top z = 0 \qquad \{\text{moltiplicando per } \tfrac{1}{2}\}$$

$$\tfrac{1}{2}X^\top z + \tfrac{1}{2}X^\top 1 = \tfrac{1}{2}X^\top 1 \qquad \{\text{aggiungendo } \tfrac{1}{2}X^\top 1\}$$

The obtained formulation:

$$X^\top \underbrace{\left(\frac{1}{2}z + \frac{1}{2}1\right)}_{\eta} = \underbrace{\frac{1}{2}X^\top 1}_{b} \tag{1}$$

permits the expression of the dual problem as follows:

$$\max_{\eta}\{y^\top \eta | X^\top \eta = b, \eta \in [0,1]^n\}$$

The LP formulation of the QR problem

Although the role of $1/2$ in the equation:

$$X^\top \underbrace{\left(\frac{1}{2}z + \frac{1}{2}1\right)}_{\eta} = \underbrace{\frac{1}{2}X^\top 1}_{b} \tag{2}$$

is seemingly neutral, it is the key to the generalization of the conditional median to the conditional quantiles

Criterion for the generic conditional quantile θ^{th}

$$\min_{\beta(\theta)} \sum_{i=1}^{n} \rho_\theta(y_i - x_i^\top \beta(\theta))$$

A similar set of steps leads to the following dual formulation for the generic quantile regression problem:

$$\max_{z}\{y^\top z | X^\top z = (1-\theta)X^\top 1, z \in [0,1]^n\},$$

where $(1-\theta)$ plays the same role that $1/2$ played for the median formulation

Technical insight: a summary

- The mean and the quantiles are particular centers of a distribution minimizing a squared sum of deviations and a weighted sum of deviations, respectively
- This idea is easily generalized to the regression setting in order to estimate conditional mean and conditional quantiles
- The development and dissemination of QR started with the formulation of the QR problem as a LP problem. Such formulation allows to exploit efficient methods and algorithms to solve a complex optimization problem offering the way to explore the whole conditional distribution of a variable and not only its center
- The QR problem typically exploits a variant of the well–known simplex algorithm for a moderate size problem. In case of datasets with a large number of units and/or covariates, interior–point methods and/or heuristic approaches have introduced

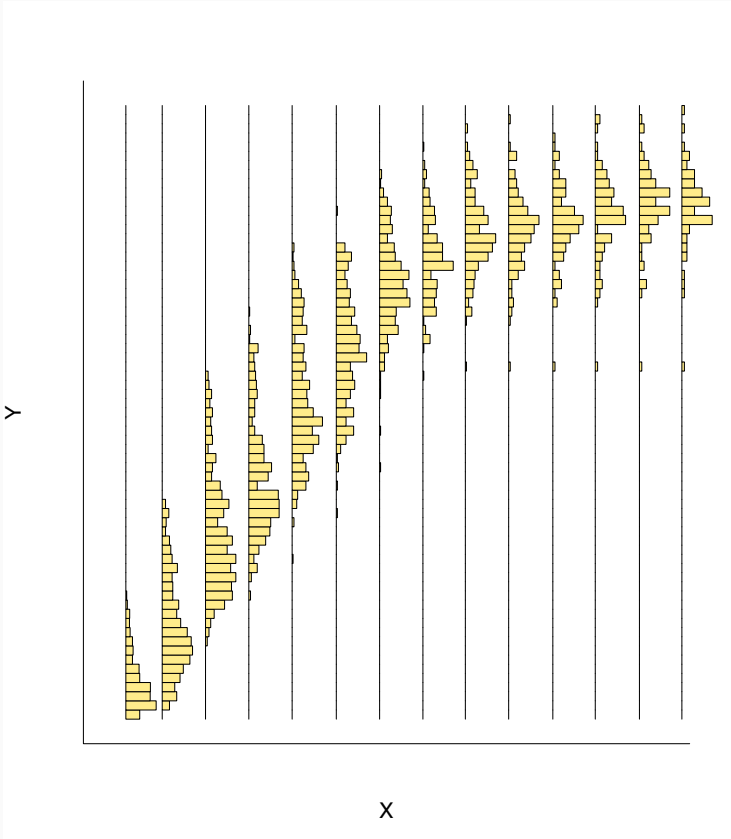
Epilogue

A more complex world

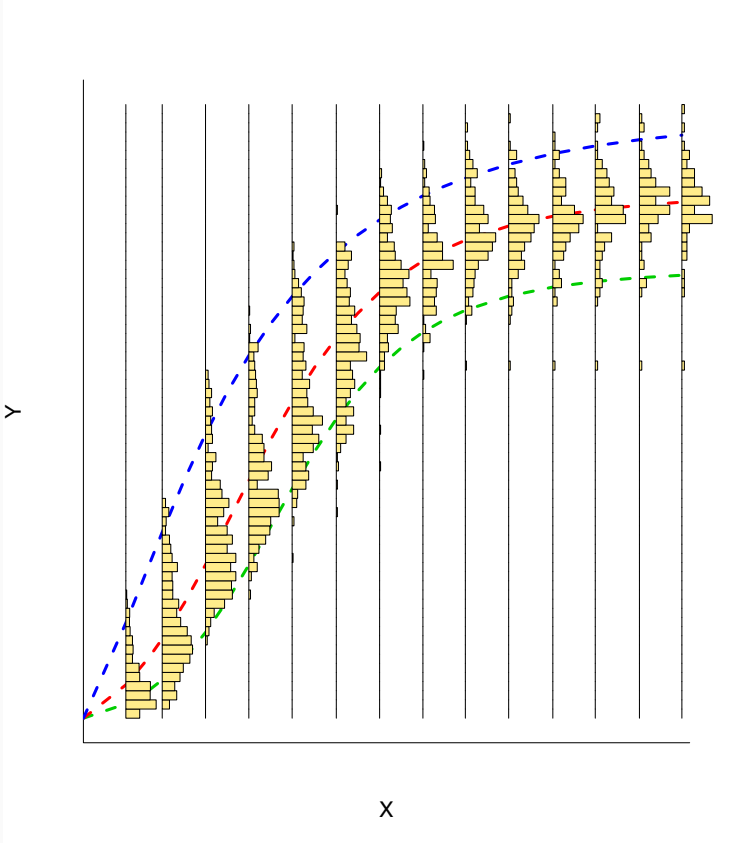
We live in a paradoxical world, where the only true safety, true though limited, comes from admitting both our uncertainty and the incompleteness with which we are able to meet it.

J. W. Tukey (1997)

A more complex world



A more complex world



Beyond the mean (epilogue)

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of X's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set.

Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

Mostseller and Tukey (1977)

Beyond the mean (epilogue)

QR offers information on the whole conditional distribution of the response variable, allowing us to discern effects that would otherwise be judged equivalent using only conditional expectation.

Nonetheless, the QR ability to statistically detect more effects can not be considered a panacea for investigating relationships between variables: in fact, the improved ability to detect a multitude of effects forces the investigator to clearly articulate what is important to the process being studied and why.