ECAS Summer Course

Introduction to Quantile Regression

Roger Koenker
University of Illinois at Urbana-Champaign

La Roche-en-Ardennes: September, 2005

My thanks goes to many collaborators: Gib Bassett, Steve Portnoy, Pin Ng, Jana Jurečková, Jose Machado, Zhijie Xiao, Lingjie Ma, Gregory Kordas and Ivan Mizera, to the US National Science Foundation for continuing support, and to the open source software community for making many of the tools of my trade. The projected version of these slides were produced by ppower4 and pdflatex; all computations and graphics were done in the R language.
Motivation

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.
Motivation

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of 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.
Motivation

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.

Mosteller and Tukey (1977)
Boxplot of CEO Pay by Firm Size

firm market value in billions

annual compensation in millions

● ● ● ●

0.1 1 10 100

* *

* * * * *

+ + + + +

+ + +

+ + + + +

+ + +

+ + + + +

+ + +

+ + + + +

+ + +

+ + + + +

+ + +
An Outline

• An Historical Introduction to Regression

• What is Quantile Regression?

• Two Artificial Examples

• Three Introductory Empirical Examples
  ✮ The Classical Engel Curve
  ✮ A Model of Infant Birthweight
  ✮ Maximum Daily Temperature in Melbourne
Boscovich’s Problem

Boscovich (1757) proposed estimating the ellipticity of the earth by solving:

\[
\min \sum_{i=1}^{n} |y_i - \alpha - x_i'\beta|
\]

subject to:

\[
\sum_{i=1}^{n} (y_i - \alpha - x_i'\beta) = 0,
\]

The constraint allowed the problem to be reduced to the weighted median problem:

\[
\min \sum_{i=1}^{n} |\tilde{y}_i - \tilde{x}_i'\beta|
\]

where \(\tilde{y}\) and \(\tilde{x}\) denote deviations from their means.
Laplace’s *Methode de Situation*

Laplace showed that Boscovich’s problem:

\[
\min \sum_{i=1}^{n} |\tilde{y}_i - \tilde{x}_i' \beta|
\]

could be solved by ordering the candidate slopes \(\{\tilde{y}_i/\tilde{x}_i\}\) and finding the weighted median using weights \(w_i = |\tilde{x}_i|\), i.e. finding the smallest \(j\) such that,

\[
\sum_{i=1}^{j} w(i) > \frac{1}{2} \sum_{i=1}^{n} w(i). 
\]
An Historical Prelude: The Ellipticity of the Earth

Figure 1: Grapefruit or Lemon? Boscovich (1755) considered the five measurements depicted above. Arc length is measured in toise per degree of latitude minus 56,700. An upward slope in this figure indicates that the earth is oblate (like a grapefruit) rather than prolate (like a lemon).
Figure 2: Boscovich (1755) computed all the pairwise slopes and initially reported a trimmed mean of the pairwise slopes as a point estimate of the earth’s ellipticity.
Figure 3: In 1757 Boscovich reestimated the model minimizing the sum of absolute residuals subject to the constraint that the residuals summed to zero to obtain the solid line fitting the triangular points. Modern measurements of the arc lengths are represented by the circles.
Earth’s Ellipticity: A Quantile Regression View

Figure 4: The quantile regression analysis of the Boscovich data identifies four distinct pairs of points that solve the weighted $\ell_1$ problem for various intervals of the parameter $\tau$. 
Sample Quantiles via Optimization

The \( \tau \)th sample quantile can be defined as any solution to:

\[
\hat{\alpha}(\tau) = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - a)
\]

where \( \rho_\tau(u) = (\tau - I(u < 0))u \) as illustrated below.
The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m E(Y - m)^2$$
The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m E(Y - m)^2$$

The conditional mean $\mu(x) = E(Y|X = x)$ solves

$$\mu(x) = \min_m E_{Y|X=x}(Y - m(X))^2.$$
The Least Squares Meta-Model

The unconditional mean solves

\[ \mu = \min_m E(Y - m)^2 \]

The conditional mean \( \mu(x) = E(Y|X = x) \) solves

\[ \mu(x) = \min_m E_{Y|X=x} (Y - m(X))^2. \]

Similarly, the unconditional \( \tau \)th quantile solves

\[ \alpha_\tau = \min_a E\rho_\tau(Y - a) \]
The Least Squares Meta-Model

The unconditional mean solves

\[ \mu = \min_m E(Y - m)^2 \]

The conditional mean \( \mu(x) = E(Y|X = x) \) solves

\[ \mu(x) = \min_m E_{Y|X=x}(Y - m(X))^2. \]

Similarly, the unconditional \( \tau \)th quantile solves

\[ \alpha_\tau = \min_a E \rho_\tau(Y - a) \]

and the conditional \( \tau \)th quantile solves

\[ \alpha_\tau(x) = \min_a E_{Y|X=x} \rho_\tau(Y - a(X)) \]
Regression Quantiles via Optimization

The sample analogue of the foregoing population concepts yields, the nonparametric quantile regression estimator

\[ \hat{\alpha}_\tau(x) = \arg\min_{a \in A} \sum_{i=1}^{n} \rho_\tau(y_i - a(x_i)) \]

If we take \( A = \{ a : \mathbb{R}^p \to \mathbb{R} | a(x) = x^\top \beta, \beta \in \mathbb{R}^p \} \), then we have the linear (in parameters) quantile regression problem:

\[ \hat{\beta}(\tau) = \arg\min_{b \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b) \]
Quantile Regression: The Movie

- Bivariate linear model with iid Student t errors
- Conditional quantile functions are parallel in green
- 100 observations indicated in blue
- Fitted quantile regression lines in red
Quantile Regression in the iid Error Model

[ 0.078, 0.085 ]
Quantile Regression in the iid Error Model

\[ [0.147, 0.165] \]
Quantile Regression in the iid Error Model

[0.184, 0.211]
Quantile Regression in the iid Error Model

[ 0.275 , 0.288 ]
Quantile Regression in the iid Error Model

\[ [0.353, 0.372] \]
Quantile Regression in the iid Error Model

[0.441, 0.444]
Quantile Regression in the iid Error Model

\[ [0.509, 0.528] \]
Quantile Regression in the iid Error Model

\[ [0.595, 0.603] \]
Quantile Regression in the iid Error Model

\[ [0.675, 0.689] \]
Quantile Regression in the iid Error Model
Quantile Regression in the iid Error Model

\[
\begin{align*}
[0.924, 0.943]
\end{align*}
\]
Virtual Quantile Regression II

- Bivariate quadratic model with Heteroscedastic $\chi^2$ errors
- Conditional quantile functions drawn in green
- 100 observations indicated in blue
- Fitted quadratic quantile regression lines in red
Quantile Regression in the Heteroscedastic Error Model

\[ \text{[0.069, 0.098]} \]
Quantile Regression in the Heteroscedastic Error Model

\[
\begin{bmatrix}
0.194 \\
0.194
\end{bmatrix}
\]
Quantile Regression in the Heteroscedastic Error Model

\[
[0.263, 0.273]
\]
Quantile Regression in the Heteroscedastic Error Model

\[ [0.31, 0.319] \]
Quantile Regression in the Heteroscedastic Error Model

\[ 0.385 , 0.403 \]
Quantile Regression in the Heteroscedastic Error Model

\[
\begin{bmatrix}
0.466, 0.493
\end{bmatrix}
\]
Quantile Regression in the Heteroscedastic Error Model

[ 0.561, 0.576 ]
Quantile Regression in the Heteroscedastic Error Model

\[ \left[ 0.657, 0.657 \right] \]
Quantile Regression in the Heteroscedastic Error Model

\[ [0.749, 0.75] \]
Quantile Regression in the Heteroscedastic Error Model

[0.826, 0.831]
Quantile Regression in the Heteroscedastic Error Model
Three Applications

- Engel’s Law: A Classical Economic Example
- Infant Birthweight: A Public Health Example
- Melbourne Daily Temperature: A Time Series Example
Engel’s Food Expenditure Data

Figure 5: Engel Curves for Food: This figure plots data taken from Engel’s (1857) study of the dependence of households’ food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.
**Engel’s Food Expenditure Data**

*Figure 6:* Engel Curves for Food: This figure plots data taken from Engel’s (1857) study of the dependence of households’ food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.
A Model of Infant Birthweight


- Data: June, 1997, Detailed Natality Data of the US. Live, singleton births, with mothers recorded as either black or white, between 18-45, and residing in the U.S. Sample size: 198,377.

- Response: Infant Birthweight (in grams)

- Covariates:
  - Mother’s Education
  - Mother’s Prenatal Care
  - Mother’s Smoking
  - Mother’s Age
  - Mother’s Weight Gain
Quantile Regression Birthweight Model I

Intercept

Boy

Married

Black

Mother's Age

Mother's Age^2

High School

Some College
Quantile Regression Birthweight Model II
Marginal Effect of Mother’s Age

Age Effect at 0.1 Quantile

Age Effect at 0.25 Quantile

Age Effect at 0.75 Quantile

Age Effect at 0.9 Quantile
Marginal Effect of Mother's Weight Gain

Mother's Weight Gain 15 Lbs

Mother's Weight Gain 22 Lbs

Mother's Weight Gain 39 Lbs

Mother's Weight Gain 47 Lbs
AR(1) Model of Melbourne Daily Temperature
Figure 7: The plot illustrates 10 years of daily maximum temperature data for Melbourne, Australia as an AR(1) scatterplot. Superimposed are estimated conditional quantile functions for $\tau \in \{.05, .10, \ldots, .95\}$. 
Conditional Densities of Melbourne Daily Temperature

Yesterday’s Temp 11

Yesterday’s Temp 16

Yesterday’s Temp 21

Yesterday’s Temp 25

Yesterday’s Temp 30

Yesterday’s Temp 35
Conclusions

- Quantile regression methods complement established mean regression (least-squares) methods.
Conclusions

- Quantile regression methods complement established mean regression (least-squares) methods.

- By focusing on local slices of the conditional distribution, they offer a useful deconstruction of conditional mean models.
Conclusions

• Quantile regression methods complement established mean regression (least-squares) methods.

• By focusing on local slices of the conditional distribution, they offer a useful deconstruction of conditional mean models.

• They provide a more flexible role for covariate effects allowing them to influence location, scale \textit{and shape} of the response distribution.
Conclusions

- Quantile regression methods complement established mean regression (least-squares) methods.

- By focusing on local slices of the conditional distribution, they offer a useful deconstruction of conditional mean models.

- They provide a more flexible role for covariate effects allowing them to influence location, scale and shape of the response distribution.

- In applications a variety of unobserved heterogeneity phenomena are rendered observable.