WILEY SERIES IN PROBABILITY AND STATISTICS

Quantile Regression

Theory and Applications

Cristina Davino
Marilena Furno
Domenico Vistocco

WILEY

Quantile Regression

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART AND SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Quantile Regression

Theory and Applications

Cristina Davino

Department of Political Science, Communication and International Relations, University of Macerata, Italy

Marilena Furno

Department of Agriculture
University of Naples Federico II, Italy

Domenico Vistocco

Department of Economics and Law
University of Cassino, Italy

WILEY

This edition first published 2014

Registered Office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Davino, Cristina.
Quantile regression : theory and applications / Cristina Davino, Marilena Furno, Domenico Vistocco.
pages cm – (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 978-1-119-97528-1 (hardback)
1. Quantile regression. 2. Regression analysis. I. Furno, Marilena, 1957–. II. Vistocco, Domenico. III. Title.
QA278.2.D38 2013
519.5'36–dc23
2013023591

A catalogue record for this book is available from the British Library.

The cover image contains a detail of the cover plate of the 'Tomb of Diver', reproduced by kind permission of the Archaeological Museum of Paestum, Italy (grant n. 19/2013 Ministero per i Beni e le Attività Culturali, Soprintendenza per i Beni Archeologici di Salerno, Avellino, Benevento e Caserta, Italy).
The detail was drawn from a photo of the Museum collection (authors: Francesco Valletta e Giovanni Grippo)

ISBN: 978-1-119-97528-1

Set in 10/12pt Times by SPi Publishers Services, Pondicherry, India

1 2014

Contents

Preface

In his seminal paper 'The Future of Data Analysis', John Tukey[1] wondered:

> 'How is novelty most likely to begin and grow?'

His answer can be summarized as follows:

- 'We should seek out wholly new questions to be answered'.
- 'We need to tackle old problems in more realistic frameworks'.
- 'We should seek out unfamiliar summaries of observational material and their useful properties'.

The topics treated in this volume provide an answer to the *novelty principle* posed by Tukey. Starting from the pioneering paper of Koenker and Basset[2], research on quantile regression has exponentially grown over the years. The development of the theory, together with the wide variety of applications, attests to the maturity of the method and proves its capability to deal with real problems.

Quantile regression allows us to look beyond the average and to provide a description of the whole conditional distribution of a response variable in terms of a set of explanatory variables. It offers, therefore, an invaluable tool to discern effects that would be otherwise lost in the classical regression model analyzing the sole conditional mean: to look beyond the average wholly allows *new questions to be answered*.

The nature of quantile regression and the ability to deal with different types of distributions allows us to eliminate dependence upon the normality assumptions and *to tackle old problems in a more realistic framework*.

The wealth of information provided by the analysis of the whole conditional distribution provides a strong incentive for the researcher to *seek out unfamiliar summaries of observational material*.

With this volume, we hope to provide an additional contribution to the diffusion of quantile regression. We are confident that the opportunity to include quantile

[1] Tukey JW 1962 The future of data analysis. *The Annals of Mathematical Statistics* **33**(1), 1–67.
[2] Koenker R and Basset G 1978 Regression quantiles. *Econometrica* **46**(1).

regression in the toolkit of applied researchers will offer more possibilities to meet the last and more demanding Tukey's point:

> '. . . and still more novelty can came from finding and evading still deeper lying constraints'.

This book contains an accompanying website. Please visit www.wiley.com/go/quantile_regression

Cristina Davino, Marilena Furno and Domenico Vistocco

Acknowledgments

Writing a book is a task in many ways similar to a long journey: the initial enthusiasm and desire to fully live a new, and somewhat unique, experience come into conflict with the natural fatigue of being a long time 'away from home'.

The authors are indebted to all those who have accompanied them along this journey. In particular they wish to thank Wiley's staff, Richard Davies, Heather Kay and Jo Taylor, who offered a discreet but constant presence throughout the entire journey.

The final structure of the volume has benefited from the comments of the anonymous referees, who evaluated the initial project work: we hope to have made the most of their suggestions.

This lengthy project has benefited from the invaluable comments and suggestions from those who have worked with us during this period (in alphabetic order): Dario Bruzzese, Vincenzo Costa, Antonella Costanzo, Alfonso Iodice D'Enza, Michele La Rocca, Mario Padula, Domenico Piccolo, Giovanni C. Porzio and Xavier Vollenweider. The remaining errors and omissions are the authors' responsibility.

Finally, our gratitude goes to our families for their continuous support: they made us feel 'at home' even when we were traveling.

Introduction

Quantile regression is a topic with great potential and is very fertile in terms of possible applications. It is growing in importance and interest, as evidenced by the increasing number of related papers appearing in scientific journals. This volume is intended as a practical guide to quantile regression. It provides empirical examples along with theoretical discussions on the issues linked to this method, with applications covering different fields.

An attempt to balance formal rigor with clarity has been made. The text concentrates on concepts rather than mathematical details, meanwhile seeking to keep the presentation rigorous. The description of the methodological issues is accompanied by applications using real data.

Computer codes for the main statistical software that include quantile regression analysis (R, SAS and Stata) are provided in the appendices, while datasets are gathered on the companion website.

The book is intended for researchers and practitioners in different fields; Statistics, Economics, Social Sciences, Environments, Biometrics and Behavioral Sciences, among others. It is aimed both for self-study by people interested in quantile regression but it can also be used as a reference text for a particular course covering theory and applications of quantile regression.

Structure of the book

Chapter 1, *A visual introduction to quantile regression*, offers a visual introduction to quantile regression starting from the simplest model with a dummy predictor, and then moving to the simple regression model with a quantitative predictor, passing through the case of a model with a nominal regressor. The chapter covers the basic idea of quantile regression and its solution in terms of a minimization problem. By the end of this chapter, the reader will be able to grasp the added value offered by quantile regression in approximating the whole distribution of a response variable in terms of a set of regressors.

Chapter 2, *Quantile regression: Understanding how and why*, deals with the quantile regression problem and its solution in terms of a linear programming problem. Such formulation historically decreed the propagation of quantile regression allowing to exploit efficient methods and algorithms to compute the solutions. The chapter also discusses the quantile regression

capability of dealing with different types of error distribution, introducing its behavior in the case of regression models characterized by homogeneous, heterogeneous and dependent error models.

Chapter 3, *Estimated coefficients and inference*, enters into more technical details. It shows the behavior of quantile regression using datasets with different characteristics. In particular it deals with the empirical distribution of the quantile regression estimator in the case of independent and identically distributed (i.i.d.) errors, non-identically distributed errors and dependent errors. The chapter then analyzes only the case of i.i.d. errors, while the other two cases are deferred to Chapter 5. The tests to verify hypotheses on more than one coefficient at a time are introduced to evaluate the validity of the selected explanatory variables.

Chapter 4, *Additional tools for the interpretation and evaluation of the quantile regression model*, discusses some typical issues arising from real data analysis. It offers keys to properly analyze data, to interpret and describe the results and to validate the model. Moreover, the effect of variable centring and scaling on the interpretation of the results is explored, both from a descriptive and from an inferential point of view. The estimation of the conditional density of the response variable and the peculiarities of the main bootstrap methods are also considered.

Chapter 5, *Models with dependent and with non-identically distributed data*, focuses on the quantile regression estimators for models characterized by heteroskedastic and by dependent errors. In particular it considers the precision of the quantile regression model in the case of i.i.d. errors, taking a closer look at the computation of confidence intervals and hypothesis testing on each estimated coefficient. It extends the analysis to the case of non-identically distributed errors, discussing different ways to verify the presence of heteroskedasticity in the data and it takes into account the case of dependent observations, discussing the estimation process in the case of a regression model with serially correlated errors.

Chapter 6, *Additional models*, where several real datasets are used to show the capabilities of some more advanced quantile regression models. In particular the chapter deals with some of the main extensions of quantile regression: its application in nonparametric models and nonlinear relationships among the variables, in the presence of censored and longitudinal data, when data are derived from different groups and with dichotomous dependent variables.

Appendices A, B and C, *Quantile regression analysis and surroundings using R, SAS and Stata*, show the commands useful to exploit a data analysis in the *R* environment (Appendix A), in *SAS* (Appendix B) and in *Stata* (Appendix C). Such appendices, far from being exhaustive, provide a description of the codes needed to load data, to visually and numerically explore the variables contained in the dataset, and to compute quantile regressions. Some commands for exporting results are briefly discussed.

A very short course would cover the linear quantile regression model of Chapter 1, data pre-processing and model validation considered in Chapter 4, and treatment of autoregressive and heteroskedastic errors of Chapter 5. Chapter 2, on the linear programming method and on the behavior of quantile regression in case of different types of error distributions, Chapter 3, on the behavior of the quantile regression estimator in the i.i.d., non-identically distributed and dependent errors, and Chapter 6, dealing with generalizations of the quantile regression to censored data, can be postponed by readers.

Although the whole project is shared by the three authors, they contributed separately to the various parts of the book: Cristina Davino developed Chapters 4 and 6 (Section 6.5 jointly with Domenico Vistocco), Marilena Furno wrote Chapters 3 and 5, Domenico Vistocco developed Chapters 1 and 2 and the three appendices on the software (Appendix B jointly with Cristina Davino and Appendix C jointly with Marilena Furno).

Nomenclature

- Vectors: $\mathbf{x}$ (lower case bold letters). The subscript $[n]$ denotes the vector dimension where the notation $\mathbf{x}_{[n]}$ is used.
- Matrices: $\mathbf{X}$ (upper case bold letters). The subscript $[n \times p]$ denotes the matrix dimensions where the notation $\mathbf{X}_{[n \times p]}$ is used.
- Transpose operator: $\top$ (e.g., $\mathbf{x}^\top$)
- Random variable: X
- Cumulative distribution function: $F_Y(y)$, where the Y subscript denotes the variables on which the function is computed. The shortened notation $F(y)$ is used where there is no risk of ambiguity.
- Quantile function: $Q_Y(\theta)$, where the Y subscript denotes the variables on which the quantile is computed. The shortened notation $Q(\theta)$ is used where there is no risk of ambiguity.
- i-th vector element: x_i
- i-th matrix row: $\mathbf{x}_i$
- Null vector: $\mathbf{0}$
- Identity vector: $\mathbf{1}$
- Identity matrix: $\mathbf{I}$
- Sample size: n
- Number of regressors: p
- Quantile: θ
- Number of estimated quantiles: k
- Quantile regression parameter: $\beta(\theta)$
- Quantile regression estimate: $\hat{\beta}(\theta)$

- Simple quantile regression model: $Q_\theta(\mathbf{y}|\mathbf{x}) = \mathbf{x}\beta(\theta) + \mathbf{e}$
- Multiple quantile regression model: $Q_\theta(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta(\theta) + \mathbf{e}$
- Loss or check function: $\rho_\theta(\mathbf{y})$
- Simple regression model: $\mathbf{y} = \beta_0 + \beta_1\mathbf{x} + \mathbf{e}$

1

A visual introduction to quantile regression

Introduction

Quantile regression is a statistical analysis able to detect more effects than conventional procedures: it does not restrict attention to the conditional mean and therefore it permits to approximate the whole conditional distribution of a response variable.

This chapter will offer a visual introduction to quantile regression starting from the simplest model with a dummy predictor, moving then to the simple regression model with a quantitative predictor, through the case of a model with a nominal regressor.

The basic idea behind quantile regression and the essential notation will be discussed in the following sections.

1.1 The essential toolkit

Classical regression focuses on the expectation of a variable Y conditional on the values of a set of variables $\mathbf{X}$, $E(Y|\mathbf{X})$, the so-called regression function (Gujarati 2003; Weisberg 2005). Such a function can be more or less complex, but it restricts exclusively on a specific location of the Y conditional distribution. Quantile regression (QR) extends this approach, allowing one to study the conditional distribution of Y on $\mathbf{X}$ at different locations and thus offering a global view on the interrelations between Y and $\mathbf{X}$. Using an analogy, we can say that for regression problems, QR is to classical regression what quantiles are to mean in terms of describing locations of a distribution.

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

QR was introduced by Koenker and Basset (1978) as an extension of classical least squares estimation of conditional mean models to conditional quantile functions. The development of QR, as Koenker (2001) later attests, starts with the idea of formulating the estimation of conditional quantile functions as an optimization problem, an idea that affords QR to use mathematical tools commonly used for the conditional mean function.

Most of the examples presented in this chapter refer to the *Cars93* dataset, which contains information on the sales of cars in the USA in 1993, and it is part of the *MASS* R package (Venables and Ripley 2002). A detailed description of the dataset is provided in Lock (1993).

1.1.1 Unconditional mean, unconditional quantiles and surroundings

In order to set off on the QR journey, a good starting point is the comparison of mean and quantiles, taking into account their objective functions. In fact, QR generalizes univariate quantiles for conditional distribution.

The comparison between mean and median as centers of an univariate distribution is almost standard and is generally used to define skewness. Let Y be a generic random variable: its mean is defined as the center c of the distribution which minimizes the squared sum of deviations; that is as the solution to the following minimization problem:

$$\mu = \underset{c}{\operatorname{argmin}}\, E(Y - c)^2. \tag{1.1}$$

The median, instead, minimizes the absolute sum of deviations. In terms of a minimization problem, the median is thus:

$$Me = \underset{c}{\operatorname{argmin}}\, E|Y - c|. \tag{1.2}$$

Using the sample observations, we can obtain the sample estimators $\hat{\mu}$ and $\hat{Me}$ for such centers.

It is well known that the univariate quantiles are defined as particular locations of the distribution, that is the θ-th quantile is the value y such that $P(Y \leq y) = \theta$. Starting from the cumulative distribution function (CDF):

$$F_Y(y) = F(y) = P(Y \leq y), \tag{1.3}$$

the quantile function is defined as its inverse:

$$Q_Y(\theta) = Q(\theta) = F_Y^{-1}(\theta) = \inf\{y : F(y) > \theta\} \tag{1.4}$$

for $\theta \in [0, 1]$. If $F(.)$ is strictly increasing and continuous, then $F^{-1}(\theta)$ is the unique real number y such that $F(y) = \theta$ (Gilchrist 2000). Figure 1.1 depicts the empirical CDF [Figure 1.1(a)] and its inverse, the empirical quantile function [Figure 1.1(b)],

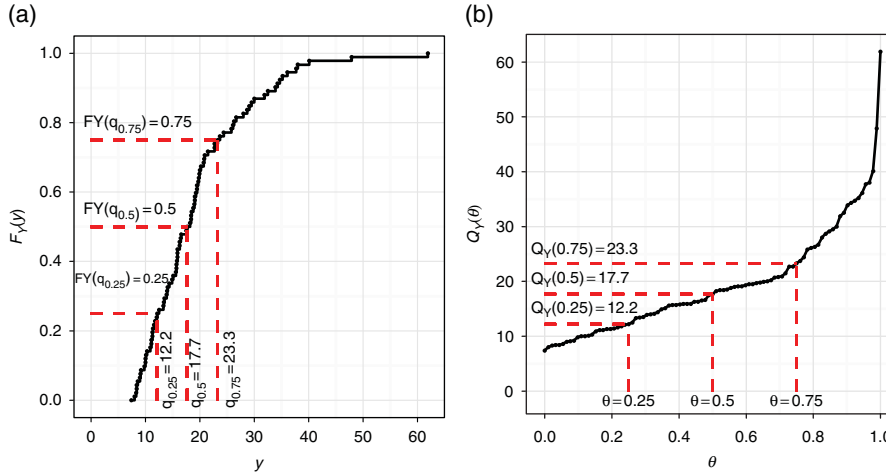

Figure 1.1 Empirical distribution function (a) and its inverse, the empirical quantile function (b), for the Price *variable of the* Cars93 *dataset. The three quartiles of* Price *are represented on the two plots: q_θ corresponds to the abscissa on the* $F_Y(y)$ *plot, while it corresponds to the ordinate on the* $Q_Y(\theta)$ *plot; the other input being the value of* θ.

for the *Price* variable of the *Cars93* dataset. The three quartiles, $\theta = \{0.25, 0.5, 0.75\}$, represented on both plots point out the strict link between the two functions.

Less common is the presentation of quantiles as particular centers of the distribution, minimizing the weighted absolute sum of deviations (Hao and Naiman 2007). In such a view the θ-th quantile is thus:

$$q_\theta = \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y - c)] \tag{1.5}$$

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

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

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

In the case of a discrete variable Y with probability distribution $f(y) = P(Y = y)$, the previous minimization problem becomes:

$$\begin{aligned}q_\theta &= \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y - c)] \\ &= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \sum_{y \leq c} |y - c| f(y) + \theta \sum_{y > c} |y - c| f(y) \right\}.\end{aligned}$$

The same criterion is adopted in the case of a continuous random variable substituting summation with integrals:

$$\begin{aligned} q_\theta &= \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y-c)] \\ &= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \int_{-\infty}^{c} |y-c| f(y) d(y) + \theta \int_{c}^{+\infty} |y-c| f(y) d(y) \right\} \end{aligned}$$

where $f(y)$ denotes the probability density function of Y. The sample estimator $\hat{q}_\theta$ for $\theta \in [0, 1]$ is likewise obtained using the sample information in the previous formula. Finally, it is straightforward to say that for $\theta = 0.5$ we obtain the median solution defined in Equation (1.2).

A graphical representation of these concepts is shown in Figure 1.2, where, for the subset of *small* cars according to the *Type* variable, the mean and the three quartiles for the *Price* variable of the *Cars93* dataset are represented on the x-axis, along with the original data. The different objective function for the mean and the three quartiles are shown on the y-axis. The quadratic shape of the mean objective function is opposed to the V-shaped objective functions for the three quartiles, symmetric for the median case and asymmetric (and opposite) for the case of the two extreme quartiles.

1.1.2 Technical insight: Quantiles as solutions of a minimization problem

In order to show the formulation of univariate quantiles as solutions of the minimization problem (Koenker 2005) specified by Equation (1.5), the presentation of the solution for the median case, Equation (1.2), is a good starting point. Assuming, without loss of generality, that Y is a continuous random variable, the expected value of the absolute sum of deviations from a given center c can be split into the following two terms:

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

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 to the solution for the minimum:

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

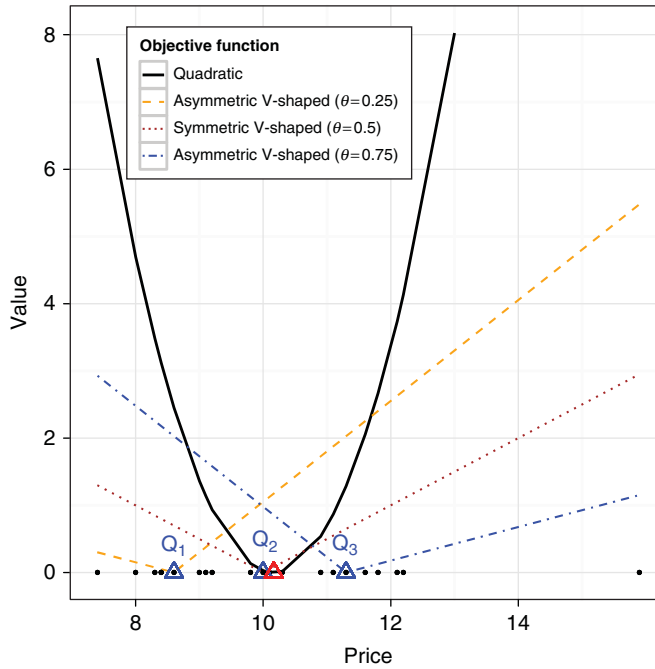

Figure 1.2 Comparison of mean and quartiles as location indexes of a univariate distribution. Data refer to the Price *of* small *cars as defined by the* Type *variable (*Cars93 *dataset). The car prices are represented using dots on the* x*-axis while the positions of the mean and of the three quartiles are depicted using triangles. Objective functions associated with the three measures are shown on the* y*-axis. From this figure, it is evident that the mean objective function has a quadratic shape while the quartile objective functions are V-shaped; moreover it is symmetric for the median case and asymmetric in the case of the two extreme quartiles.*

The solution can then be obtained applying the derivative and integrating per part as follows:

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

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

Taking into account that:

$$\lim_{x\to-\infty} f(x) = \lim_{x\to+\infty} f(x) = 0$$

for a well-defined probability density function, the integrand restricts in $y = c$[1]:

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

Using then the CDF definition, Equation (1.3), the previous equation reduces to:

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

and thus:

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

The solution of the minimization problem formulated in Equation (1.2) is thus the median. The above solution does not change by multiplying the two components of $E|Y - c|$ by a constant θ and $(1 - \theta)$, respectively. This allows us to formulate the same problem for the generic quantile θ. Namely, using the same strategy for Equation (1.5), we obtain:

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

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:

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

1.1.3 Conditional mean, conditional quantiles and surroundings

By replacing the sorting with optimization, the above line of reasoning generalizes easily to the regression setting. In fact, interpreting Y as a response variable and $\mathbf{X}$ as a set of predictor variables, the idea of the unconditional mean as the minimizer of Equation (1.1) can be extended to the estimation of the conditional mean function:

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

[1] It is worth to recall that our interest is in $y = c$ because it is where $E|Y - c|$ is minimized.

where $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = E[Y|\mathbf{X} = \mathbf{x}_i]$ is the conditional mean function. In the case of a linear mean function, $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$, the previous equation becomes:

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

yielding the least squares linear regression model. The problem of minimizing the squared error can then be reduced to a problem of numerical linear algebra. Using again the *Cars93* dataset, Figure 1.3(a) shows the geometric interpretation of the least squares criterion in the case of a simple linear regression model where *Price* is the response variable and *Horsepower* is the predictor. The least squares solution provides the line that minimizes the sum of the area of the squares as determined by

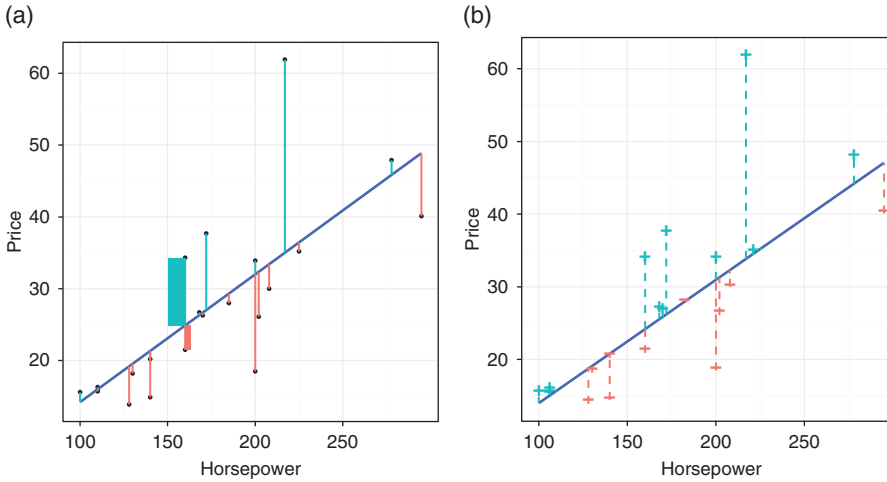

Figure 1.3 A geometric comparison of least squares (a) and least absolute (b) criteria for a simple linear regression problem. Data refer to the relationship between Price *(response variable) and* Horsepower *(predictor variable) for the* Midsize *subset of cars, as determined by the* Type *variable. Minimization of the squared errors, as required by the least squares criterion (a), is geometrically equivalent to minimizing the sum of the squares obtained by projecting the points on the regression line perpendicularly to the* x*-axis. Two of these squares, one corresponding to a negative deviation and one for a positive deviation, are shown on the plot. It is worth noticing that the choice of using non-equal scales on both the axes involves rectangles and not squares in an initial analysis of the chart. (b) shows the least absolute criterion for the median case: the minimization of the sum of the least absolute deviations from the conditional median is equivalent to the overall minimization of the lengths of the segments obtained by the* x*-axis perpendicular projection of the points. Negative* $(-)$ *and positive* $(+)$ *deviations share the same weight in determining the line, involving then a symmetric loss function. As the number of points is even, there is no point that lies on the line.*

the projection of the observed data on the same line using segments perpendicular to the x-axis. Figure 1.3(a), in particular, shows the contribution of two points to this sum of squares, one point lying below the line and one point lying above the line.

The same approach can be used to extend Equation (1.2) to the median, or the more general Equation (1.5) to the generic θ-th quantile. In this latter case, we obtain:

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

where $Q_Y(\theta, \mathbf{X}) = Q_\theta[Y|\mathbf{X} = \mathbf{x}]$ denotes the generic conditional quantile function. Likewise, for the linear model case the previous equation becomes:

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

where the (θ)-notation denotes that the parameters and the corresponding estimators are for a specific quantile θ. Figure 1.3(b) shows the geometric interpretation of this least absolute deviation criterion in the case of the model:

$$\widehat{Price} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) Horsepower,\ \theta = 0.5.$$

The solution to this median case is the line that minimizes the sum of absolute deviations, that is the sum of the lengths of the segments projecting each y_i onto the line perpendicularly to the x-axis.

Figure 1.4 shows the geometric interpretation of the QR solution for the case of $\theta = 0.25$ (first quartile) in Figure 1.4(a), and of $\theta = 0.75$ (third quartile) in Figure 1.4(b), respectively. While in the median case, the deviations $y_i \leq \hat{y}_i$ and the deviations $y_i > \hat{y}_i$ give the same contribution to the criterion, in the first quartile and third quartile cases they bring an asymmetric contribution: for $\theta = 0.25$, the deviations of $y_i \leq \hat{y}_i$ have weight $1 - \theta = 0.75$ with respect to the deviations corresponding to $y_i > \hat{y}_i$, whose weights are $\theta = 0.25$, with m points lying exactly on the line, where m is equal to the number of parameters in the model. This asymmetric weighting system involves an attraction of the line towards the negative deviation. The same happens, inverting the weights and then the direction of attraction, for the case of $\theta = 0.75$. The mathematical formulation of the problem leads to the solution of a linear programming problem. Its basic structure, and the counterpart algorithm solution, will be presented in the next chapter (see Sections 2.1.2 and 2.1.3).

1.2 The simplest QR model: The case of the dummy regressor

In order to introduce quantile regression, it is useful to begin by illustrating the simplest form of linear model; that is a model with a quantitative response variable and a dummy predictor variable. This simple setting aims to study differences in the response variable between the two groups as determined by the dichotomous predictor variable.

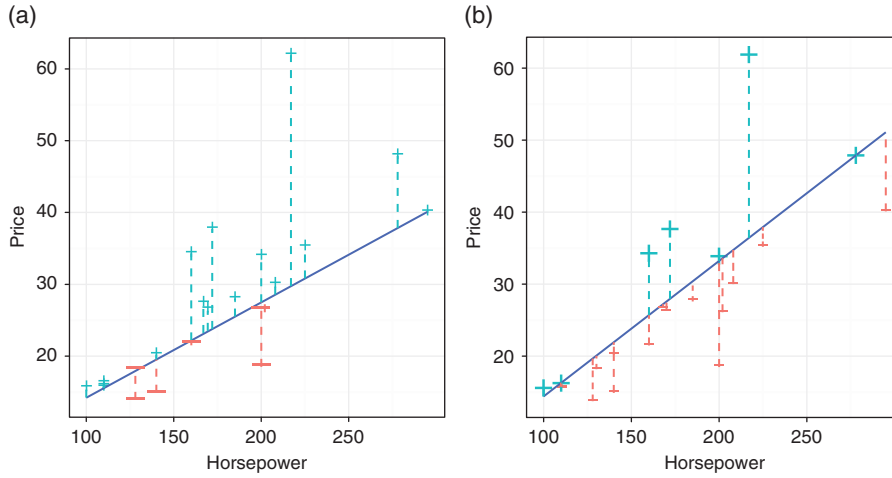

Figure 1.4 The geometric interpretation of the asymmetric least absolute deviation criterion in the case of the conditional first quartile ($\theta = 0.25$)*, in (a), and of the conditional third quartile* ($\theta = 0.75$)*, in (b). The use of an unbalanced weighting system (weight equal to* $(1-\theta)$ *for the sum of negative deviations and weight equal to* θ *for the sum of positive deviations) determines, in the two cases, a different attraction of points lying below or above the regression line. Such different attraction of the two subsets of points is depicted using sizes proportional to the corresponding subset weight. The case of an asymmetric loss function, thus involves a line separating points with almost* θ*% lying below the line and the remainder* $(1-\theta)$*% lying above the line. The explanation of this inaccurate partition is essentially due to the linear programming solution of the problem and is given in detail in Chapter 2. It is also worth noticing that* $m = 2$ *points lies exactly on the line, where* m *is equal to the number of model parameters.*

To illustrate this simple model, we refer again to the *Cars93* dataset in order to compare the distribution of the *Price* of the cars between the two groups of USA and non-USA company manufacturers. In particular, the dummy variable *Origin*, which assumes a unit value for non-USA cars and zero for USA cars, is used as regressor.

Figure 1.5 shows the dotplots for the two groups, USA cars represented on the left-hand side and non-USA cars on the right-hand side. *Price* means for the two distributions are depicted by asterisks while the three quartiles are shown using a box bounded on the first and third quartiles, the box sliced on the median. From the analysis of the two groups' dotplots in Figure 1.5, it is evident that the two samples share a similar mean, but the long right tail for the *non-USA* car distribution gives rise to strong differences in the two extreme quartiles. A different view of the difference between the two distributions is offered by the plot of the *Price* density for the two samples, shown in Figure 1.6(a), which shows a right heavy tail for the non-USA cars distribution. A third graphical representation frequently used to

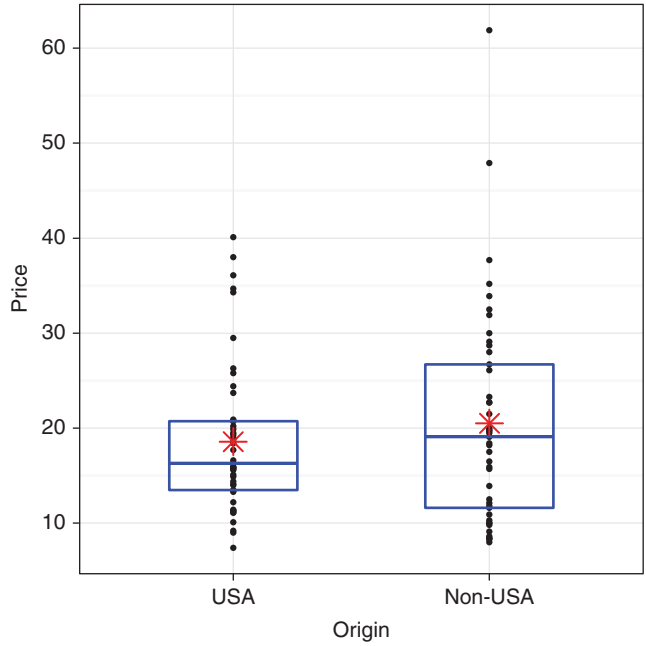

Figure 1.5 Comparison of the distribution of the Price *variable for the USA and non-USA manufactured cars, the two groups being specified by the* Origin *variable (*Cars93 *dataset). On each dotplot, an asterisk for the mean is superimposed. The two boxes are bounded on the first and third quartiles of each distribution and sliced on the corresponding medians. The right boxplot shows a larger right tail for the non-USA subset.*

compare two datasets is the Q–Q plot (Das and Bhattacharjee 2008), which represents the quantiles of the first dataset on the x-axis versus the quantiles of the second dataset on the y-axis, along with a 45^{o} reference line. If the two datasets share a common distribution, the points should fall approximately along the reference line, if the points fall under (above) the line the corresponding set shows a shorter (longer) tail in that part of the distribution. Figure 1.6(b) shows the Q–Q plot for the *Price* of the cars comparing their origin. The representation offers the same information as the density plot, allowing to evaluate shifts in location and in scale, the presence of outliers and differences in tail behavior. Unlike the density plot, which requires one to set the kernel width, the Q–Q plot does not require any tuning parameter. Moreover, it will turn out to be an enlightening representation for the QR output in the case of a simple model consisting of a dummy predictor variable.

It is well known that in this simple case of a model with a unique dummy predictor variable, the classical least square regression:

$$\widehat{Price} = \hat{\beta}_0 + \hat{\beta}_1 Origin$$

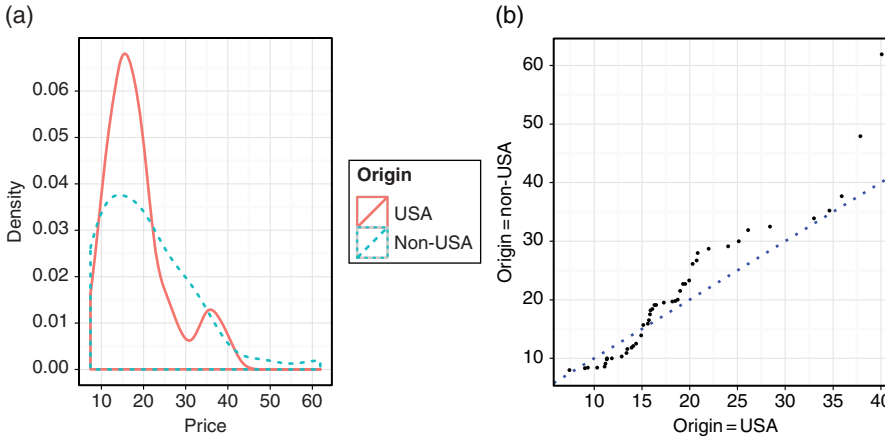

Figure 1.6 Two alternatives to boxplots for comparing distribution: density plot (a) and Q–Q plot (b). Again for the Cars93 *dataset, the two plots refer to the comparison of the* Price *variable for the two subsets as determined by the country manufacturer (USA vs non-USA origin). As well as the dotplots in Figure 1.5, both the density plot and Q–Q plot show a right heavy tail for the non-USA cars. The Q–Q plot in particular, is obtained by depicting quantiles of the USA subset of cars (*x*-axis) and quantiles of the non-USA subset (*y*-axis). It allows us to grasp the QR output in the case of a simple model with a dummy variable.*

is equivalent to the mean comparison between the two groups of USA and non-USA manufactured cars, providing the same results for the classical two samples t-test in the case of inference.

Likewise, the estimation of the QR model:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) Origin \tag{1.6}$$

for different values of $\theta \in [0, 1]$ permits us to obtain an estimation of the *Price* quantiles for the two groups of cars. Using the coding USA = 0 and non-USA = 1 for the *Origin* indicator variable in Equation (1.6), it is straigthforward that the estimated price for the USA cars is then:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{0} = \hat{\beta}_0(\theta),$$

while for the non-USA subset the model becomes:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{1} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta).$$

$\hat{\beta}_0(\theta)$ thus provides the estimation of the conditional θ quantile of the *Price* for USA cars, while, for the non-USA cars, the conditional θ quantile is obtained through

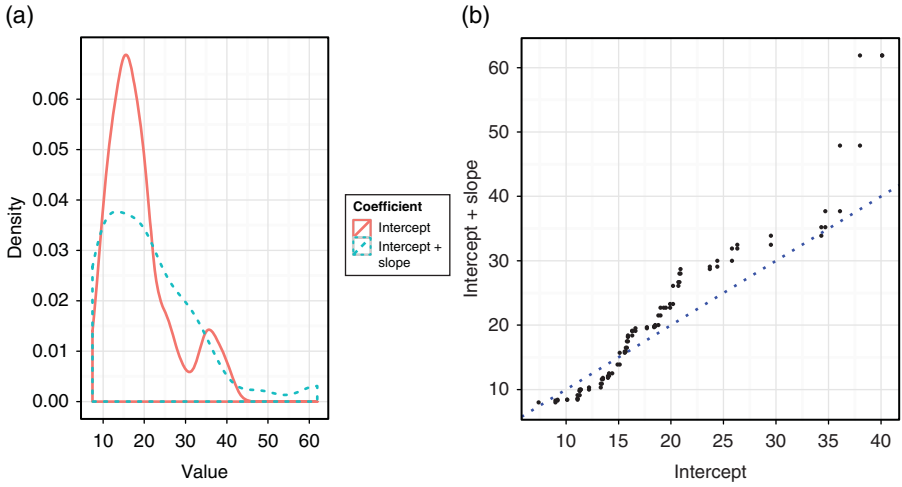

Figure 1.7 Density plot (a) and Q–Q plot (b) for the Price *variable for the two groups (USA vs non-USA) of cars, estimated through the QR model with the use of only the indicator* Origin *variable as predictor. In this simple setting, the set of intercepts for different values of* θ *estimates the* Price *distribution for the USA group while the sum of the intercepts and slopes provides the* Price *estimation for the non-USA group of cars. The QR coefficients correspond to 90 different values of* θ *in [0, 1]. The plots are practically equivalent to the observed ones (see Figure 1.6).*

the combination of the two coefficients; that is as $\hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)$. By varying θ in $[0, 1]$, the set of estimated intercepts offers an estimate of the *Price* distribution for the USA cars. For the non-USA cars, price is obtained through the sum of the sets of intercepts and slopes for the different θ. For this example, using 90 different equally spaced quantiles in the range [0,1], the QR intercepts and slopes are here used to estimate the *Price* distribution for the two groups of cars. The density plot and the Q–Q plot of the estimated distribution reported in Figure 1.7(a) and (b), respectively, are practically equivalent to the observed ones shown in Figure 1.6. From Figure 1.7, it is evident how the different sets of coefficients corresponding to different values of θ are able to fully describe the *Price* distribution conditional on the two levels of the *Origin* variable. In this simplified setting, such conditional distributions represent the estimation of the *Price* distribution for the two different groups of cars computed using the predictor indicator variable.

Numerical results for this example are reported in Table 1.1: the first two rows show the estimated QR coefficients for five selected quantiles, $\theta \in (0.1, 0.25, 0.5, 0.75, 0.9)$, and the third and fourth rows report the estimated quantiles for the two groups through the combination of the two regression coefficients; finally, the fifth and sixth rows show the quantiles computed on the observed data.

Table 1.1 QR coefficients (first two rows) for the simple model $\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)$ Origin. The combination of the two estimated coefficients for the different θ allows us to estimate the corresponding *Price* quantiles for the two groups of cars (third and fourth row). The last two rows show the unconditional *Price* quantiles.

	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
Parameter estimates					
(Intercept)	11.1	13.4	16.3	20.8	34.3
Origin (non-USA)	−2.5	−1.8	2.8	5.9	−0.4
(Intercept)	11.1	13.4	16.3	20.8	34.3
Intercept + Origin (non-USA)	8.6	11.6	19.1	26.7	33.9
Unconditional *Price* quantiles					
Origin (USA)	11.1	13.5	16.3	20.7	30.9
Origin (non-USA)	8.8	11.6	19.1	26.7	33.3

1.3 A slightly more complex QR model: The case of a nominal regressor

The simple model introduced in the previous section can be slightly complicated by replacing the indicator predictor variable by a multiple level categorical variable with *g categories*. The classical regression for this setting is substantially equivalent to a mean comparison among the *g* groups as determined by the levels of the nominal regressors. Similarly, QR allows us to compare the different quantiles among the different *g* groups.

In order to show this setting, again using the *Cars93* dataset, we consider in this section the Airbags variable as a regressor to predict the car price. For such a variable, the three levels – *None*, *Driver only* and *Driver and Passenger* – correspond to the number of *Airbags* installed in the car.

It is widely known that in a regression model, to deal with a *g* level nominal variable requires the introduction of a $g - 1$ dummy variable (Scott Long 1997). Such a coding system then produces the QR model:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) I(Driver\ only) + \hat{\beta}_2(\theta) I(Driver\ and\ Passenger),$$

where $I()$ is the indicator function returning 1 if the particular unit assumes the value in parenthesis and 0 otherwise. Moreover, it is well known that the first level, the so-called reference level, is associated with the model intercept.

The previous equation provides the estimation of the conditional θ quantile of the *Price* and the coding system allows us to easily obtain the estimation for the three different levels of the regressor. For the *None* level, associated with the intercept, the model reduces to:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{0} + \hat{\beta}_2(\theta) \times \mathbf{0} = \hat{\beta}_0(\theta).$$

Table 1.2 QR coefficients (first three rows) for the simple model $\hat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)I(\textit{Driver only}) + \hat{\beta}_2(\theta)I(\textit{Driver and Passenger})$, where *Price* is the dependent variable and *Airbags* is the regressor. The three estimated coefficients are shown in the first three rows; the combination of the three estimated coefficient for the different θ allows us to estimate the corresponding *Price* quantiles for the two groups of cars (rows four to six); the last three rows show the unconditional *Price* quantiles.

	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
(Intercept)	8.4	9.2	12.2	16.3	19.5
Airbags (*Driver only*)	3.4	6.4	7.7	10.0	12.4
Airbags (*Driver and Passenger*)	7.4	8.5	12.2	18.9	20.6
(Intercept)	8.4	9.2	12.2	16.3	19.5
Intercept + Airbags (*Driver only*)	11.8	15.6	19.9	26.3	31.9
Intercept + Airbags (*Driver and Passenger*)	15.8	17.7	24.4	35.2	40.1
Unconditional *Price* quantiles					
Airbags (*None*)	8.4	9.4	11.9	16.2	19.4
Airbags (*Driver only*)	11.9	15.6	19.9	26.2	31.5
Airbags (*Driver and Passenger*)	16.6	18.2	25.5	35.4	38.9

A similar line of reasoning leads to the estimation of the conditional quantiles for the *Drivers only* group:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{1} + \hat{\beta}_2(\theta) \times \mathbf{0} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta),$$

and, for the *Driver and Passenger* group:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{0} + \hat{\beta}_2(\theta) \times \mathbf{1} = \hat{\beta}_0(\theta) + \hat{\beta}_2(\theta).$$

Therefore, in a model with a nominal regressor, for a given quantile θ, the combination of the intercept with the different slopes, allows us to estimate the conditional quantile of the response variable. The estimated effect of the particular group is obtained using the dummy variable associated with the particular slope.

Numerical results for the simple model previously considered are in Table 1.2: the three coefficients are on the first three rows of the table; rows four to six show the estimated conditional quantile for the *Price* variable, while the last three rows show the unconditional *Price* quantiles. Using again 90 quantiles in $(0, 1)$, the estimated conditional densities of the *Price* variable for the three groups of cars are represented

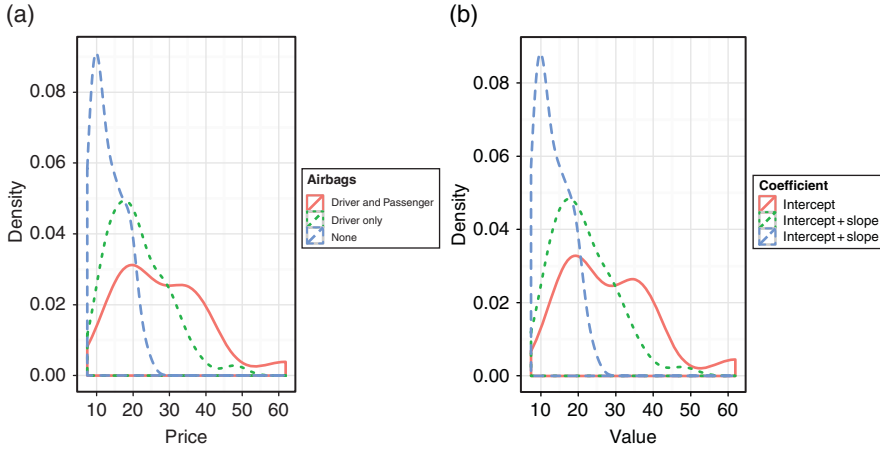

Figure 1.8 Observed density plot (a) for the Price *variable: different lines refer to different groups determined by the nominal variable,* Airbags*. Through the combination of the different coefficients it is possibile to estimate the quantile of the* Price *variable conditional on the* Airbags *variable (b). Comparing (a) and (b), it is evident that the estimated densities are practically equivalent to the observed ones.*

in Figure 1.8(b). They are practically equivalent to three observed distributions, shown in Figure 1.8(a).

1.4 A typical QR model: The case of a quantitative regressor

Once that the general idea behind QR and the two models involving a nominal regressor have been illustrated, we can move to the more common regression setting with a quantitative regressor. A simple example for the *Cars93* dataset is offered by the model:

$$\widehat{Price} = \hat{\beta}_0 + \hat{\beta}_1 Passengers;$$

that is studying the car price starting from the number of passengers they are registered to carry. For the sake of illustration, we restrict our attention to the cars licensed to carry 4, 5, and 6 passengers. Figure 1.9(a) depicts the scatterplot of the *Passengers* variable (x-axis) vs the *Price* variable (y-axis). Given the nature of the *Passengers* variable (discrete variable assuming only three different values), such a scatterplot can be easily interpreted as the combination of three dotplots for the *Price* variable corresponding to the three different numbers of passengers; that is it depicts the three conditional distributions of the *Price* on the *Passengers*. From Figure 1.9(a), differences in location and variability for the three groups are evident. Such differences

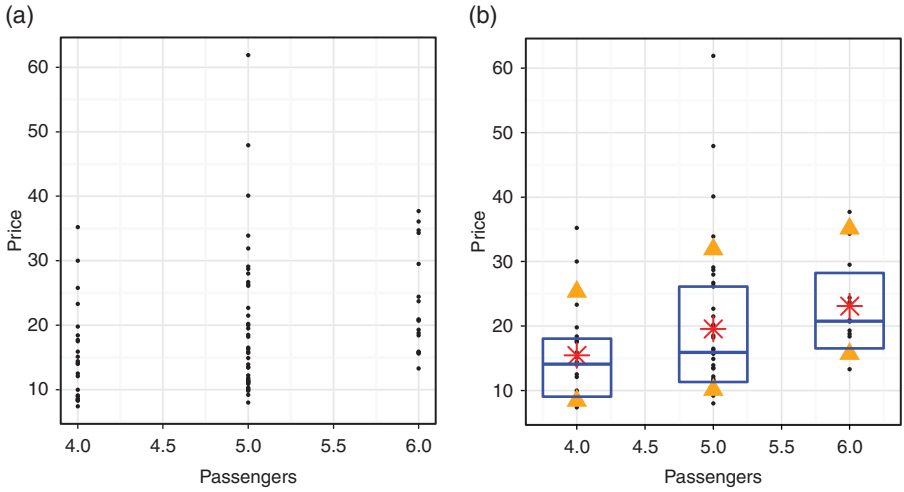

Figure 1.9 Scatterplot of Passengers *vs* Price *for the* Cars93 *dataset (a). The* Passengers *variable assumes only three different values for the considered subset: the plot can then be easily interpreted as the distributions of the* Price *variable conditional on the values of the* Passengers *variable. In (b), the location of the mean (asterisk) and of five quantiles for the three conditional distributions are superimposed on the scatterplot. In particular, the three quartiles are depicted using a box, with boundaries on the two extreme quartiles and sliced on the median, and the two extreme quantiles (*$q_{0.1}$ *and* $q_{0.9}$*) are shown with triangles.*

become clearer by superimposing the information of the mean and of five quantiles of the distributions [Figure 1.9(b)]. The five quantiles used are the common three quartiles (depicted using the box with boundaries on the two extreme quartiles and sliced on the median) and two symmetric extreme quantiles, $q_{0.1}$ and $q_{0.9}$, represented by the triangles in Figure 1.9(b), along with the means (depicted using asterisks).

Using five such values for θ, the QR simple linear model is estimated, and the results are superimposed on Figure 1.10. The ordinary least squares (OLS) line is recognizable by looking at the position of the asterisks. The patterns in Figure 1.10 reveal differences in location, variability, and shape for the different values of the number of passengers for which the car is patented. Such differences in the relationship between the *Price* variable and the *Passengers* variable are more evident by looking at the corresponding three density plots for the *Price* variable shown in Figure 1.11, along with the marginal density plot for the *Price* variable. Each panel of Figure 1.11 refers to a conditional distribution, the latter depicting the marginal distribution: the observed *Price* variable and the QR estimated *Price* variable, conditional on the values of the number of passengers, are represented in each subplot. From this figure it is evident that the estimated QR distribution is able to fully describe the patterns of the *Price*, both for the conditional cases and for the

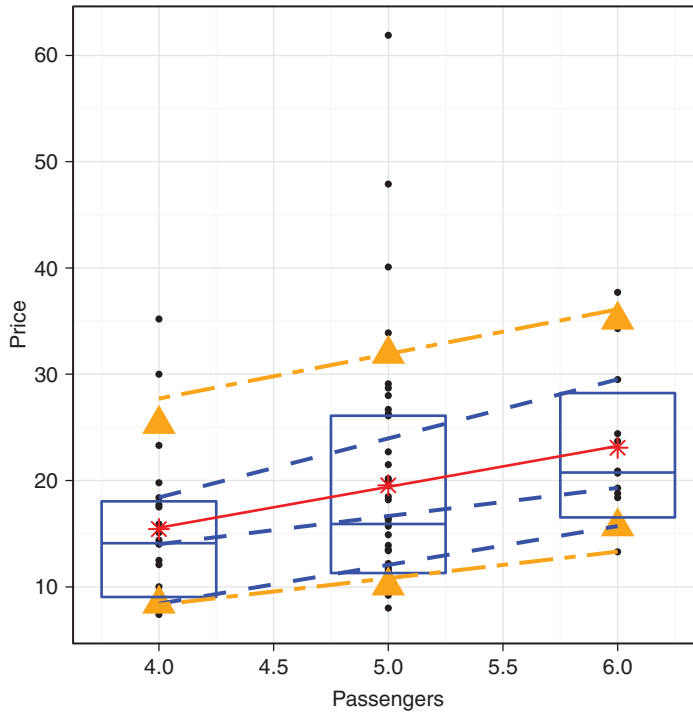

Figure 1.10 QR lines for $\theta = (0.1, 0.25, 0.5, 0.75, 0.9)$ *for the simple linear model* $\widehat{\text{Price}} = \beta_0(\theta) + \beta_1(\theta)$ Passengers. *The OLS line is recognizable by looking at the position of the asterisks, which represent the averages for the three groups.*

marginal case. Density estimation is obtained by simply combining the different estimates for different values of θ and selecting the 'best' model according to the deviation between the observed value and the estimated values. More details on the technicalities behind the density estimation are illustrated in Chapter 4, Section 4.2.

As already mentioned, QR is an extension of the classical estimation of conditional mean models to conditional quantile functions; that is an approach allowing us to estimate the conditional quantiles of the distribution of a response variable Y in function of a set $\mathbf{X}$ of predictor variables.

In the framework of a linear regression, the QR model for a given conditional quantile θ can be formulated as follows:

$$Q_\theta(Y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\theta),$$

where $0 < \theta < 1$ and $Q_\theta(.|.)$ denotes the conditional quantile function for the θ-th quantile.

The parameter estimates in QR linear models have the same interpretation as those of any other linear model, as rates of change. Therefore, in a similar way to the OLS model, the $\beta_i(\theta)$ coefficient of the QR model can be interpreted as the rate of

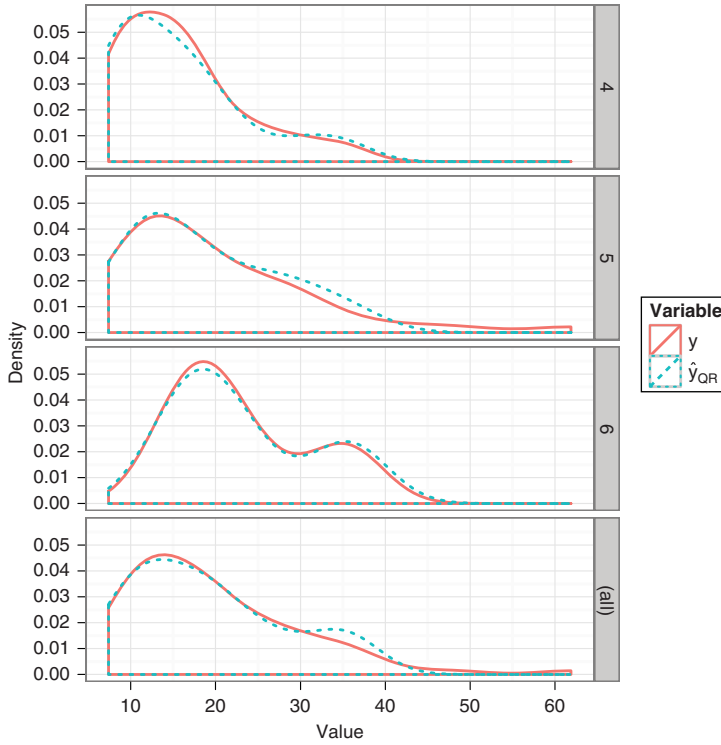

Figure 1.11 The two different lines in each panel represent the densities for the observed Price *variable and the estimated* Price *variable, the latter obtained through QR. Different panels depict the conditional distributions of the three values of the* Passengers *variable, and the bottom panel depicts the marginal distributions.*

change of the θ-th quantile of the dependent variable distribution per unit change in the value of the i-th regressor:

$$\beta_i(\theta) = \frac{\partial Q_\theta(Y|\mathbf{X})}{\partial \mathbf{x}_i}.$$

In Table 1.3, OLS and QR results obtained for the previous simple linear model are limited to the five quantiles $\theta = (0.1, 0.25, 0.5, 0.75, 0.9)$.

Table 1.3 OLS and QR coefficients for the simple linear model predicting car *Price* through the *Passengers* they are licensed to carry.

	OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
(Intercept)	0.18	−1.7	−6.20	3.40	−3.80	10.9
Passengers	3.84	2.5	3.65	2.65	5.55	4.2

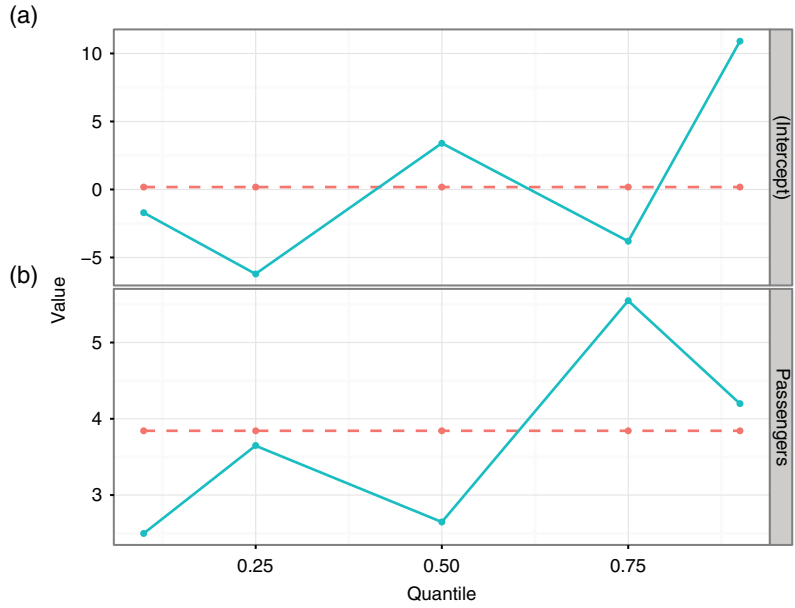

Figure 1.12 Graphical representation of the QR intercept (a) and slope (b) behavior (solid line) for the simple linear model $\widehat{\text{Price}} = \hat{\beta}_0 + \hat{\beta}_1$ Passengers. *The conditional quantiles are represented on the* x*-axis and the coefficient values on the* y*-axis. A dashed line for both coefficients is placed at the corresponding OLS estimates.*

The effect of *Passengers* increases moving from the lower part to the upper part of the *Price* distribution with the exception of the median: this confirms the conjecture of a location-scale model, namely of changes both in the central tendency and in the variability of the response variable. If in the central location (both in mean and in median) a one-unit increase of the *Passengers* variable leads to an increase in the *Price* by 3.84 and 2.65, respectively, such an effect is equal to 3.65 at the 25-th quantile and increases further to 5.55 at the 75-th quantile. In practical terms, QR results suggest that the number of passengers a car is licensed to carry has a greater effect for the more expensive cars, that is for cars with a price greater than or equal to the 75-th quantile ($\hat{\beta}_{\theta=0.75} = 5.55$).

The obtained results can also be graphically inspected. A typical graphical representation of QR coefficients (Koenker 2011) permits us to observe the different behaviors of the coefficients with respect to the different quantiles. Figure 1.12 displays the obtained estimates for intercept and slope: the different conditional quantiles are represented on the x-axis while the coefficient values are on the y-axis, the solid line with filled symbols represents the point estimates for the five distinct conditional quantiles, while the dashed line is placed at the value of the OLS estimate. With respect to the slope [Figure 1.12(b)], the previous comments on the OLS

and QR results are confirmed: up to the median, QR slope estimates are lower than the OLS counterpart, while the effect of *Passengers* on *Price* seems slightly stronger for the upper quantile. The intercept estimates seem more dependent on the particular quantile.

This chapter restricts itself to simple linear regression models in order to introduce the QR logic. It is worth noticing that the extension to multiple regression follows the same line of reasoning of classical regression (Gujarati 2003; Weisberg 2005). Using the same rules of OLS regression, a categorical explanatory variable can also be included in the model: it is preliminarily transformed into dummy variables, and all of them, except for one, are included in the model, the excluded category being the reference category in the interpretation of the results (Scott Long 1997). More realistic and interesting applications of QR, along with the inference tools, will be presented in the following chapters, starting from Chapter 3.

Finally, for all the examples shown up to now, five conditional quantiles, $\theta = (0.1, 0.25, 0.5, 0.75, 0.9)$, have been used for synthesis purposes. However, although it is possible to extract an infinite number of quantiles, in practice, a finite number is numerically distinct, the so-called *quantile process*; it depends on the number of observations increasing roughly linearly with them (Koenker and D'Orey 1987; Portnoy 1991; Buchinsky 1998). In the case of a particular interest in very high quantiles, larger samples are required (Cade and Noon 2003). For technical details about the quantile process the reader is referred to Chapter 2, Section 2.3.3.

1.5 Summary of key points

- While classical regression gives only information on the conditional expectation, quantile regression extends the viewpoint on the whole conditional distribution of the response variable.
- The mean and the quantiles are particular centers of a distribution minimizing a squared sum of deviations and a weighted absolute sum of deviations, respectively. This idea is easily generalized to the regression setting in order to estimate conditional mean and conditional quantiles.
- A simple linear regression model with a quantitative response variable and a dummy regressor allows us to compare the mean (classical regression) and the quantiles (quantile regression) between the two groups determined by the dummy regressor. Using a nominal regressor, such a comparison is among the g groups corresponding to the g levels of the nominal variable.
- For QR, as well as for classical regression, the parameter estimates in linear models are interpretable as rates of changes: the $\beta_i(\theta)$ coefficient can be interpreted as the rate of change of the θ-th quantile of the dependent variable distribution per unit change in the value of the i-th regressor.
- QR provides location, scale, and shape shift information on the conditional distribution of the response variable.

- QR allows us to approximate the whole distribution of a response variable conditional on the values of a set of regressors.

- We have seen how QR, offering information on the whole conditional distribution of the response variable, allows 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.

References

Buchinsky M 1998 Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources* **33**, 88–126.

Cade BS and Noon BR 2003 A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1**(8), 412–420.

Das KK and Bhattacharjee D 2008 *101 Graphical Techniques*. Asian Books.

Gilchrist WG 2000 *Statistical Modelling with Quantile Functions*. Chapman & Hall / CRC.

Gujarati DN 2003 *Basic Econometrics*, International Edition. McGraw-Hill.

Hao L and Naiman DQ 2007 *Quantile Regression*. SAGE Publications, Inc.

Koenker R 2001 Linear hypothesis: regression (quantile). In *International Encyclopedia of Social & Behavioral Sciences* (Smelser NJ and Baltes PB eds), 8893–8899. Elsevier.

Koenker R 2005 *Quantile Regression*. Cambridge University Press.

Koenker R 2011 *quantreg: Quantile Regression*. R package version 4.76. http://CRAN.R-project.org/package=quantreg.

Koenker R and Basset G 1978 Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker RW and D'Orey V 1987 Algorithm AS 229: computing regression quantiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **36**(3), 383–393.

Lock RH 1993 1993 New car data. *Journal of Statistics Education* **1**(1).

Portnoy S 1991 Asymptotic behavior of the number of regression quantile breakpoints. *SIAM Journal on Scientific and Statistical Computing Archive* **12**(4), 867–883.

Scott Long J 1997 *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications.

Venables WN and Ripley BD 2002 *Modern Applied Statistics with S*, 4-th Edition. Springer.

Weisberg S 2005 *Applied Linear Regression*, 3rd Edition. John Wiley & Sons, Ltd.

2

Quantile regression: Understanding how and why

Introduction

The real sounding board of a statistical method is clearly related to the possibility to test its features on real data. With such aims, the availability of efficient methods and algorithms able to compute solutions in different realistic settings is essential. The formulation of the nonlinear objective function of quantile regression through linear programming was definitely the starting point for its propagation.

With respect to the possible applications, the natural (and first) use of quantile regression to face heteroskedasticity is not the only framework in which it reveals its strength. Due to its nonparametric nature, quantile regression is a valid alternative to deal with data characterized by different types of error distribution.

After an introduction to the linear programming approach for solving the quantile regression problem, this chapter will focus on the added value of quantile regression exploring its features in the case of regression models with homogeneous, heterogeneous and dependent error models.

2.1 How and why quantile regression works

Wagner (1959) proved that the least absolute deviation criterion can be formulated as a linear programming technique and then solved efficiently exploiting proper methods and algorithms. Koenker and Basset (1978) pointed out how conditional quantiles could be estimated by an optimization function minimizing a sum of

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

weighted absolute deviations, using weights as asymmetric functions of the quantiles. The linear programming formulation of the problem was therefore natural, allowing to exploit the same methods for estimating the conditional quantiles, and offering researchers and practitioners a tool for looking inside the whole conditional distribution apart from its center.

After an introduction to linear programming features, the remainder of this section is devoted to showing the linear programming formulation of the QR problem.

2.1.1 The general linear programming problem

Linear programming is an extremely flexible tool widely applied in different fields and with different aims, such as for allocating resources, scheduling workers and plans, planning portfolio investment, formulating military strategies, and setting a marketing strategy. An exhaustive treatment of the topic would be outside the scope of this book. In this section, we present only the basic ideas useful to catch the QR solving procedure through linear programming. The interested reader may consult one of the many books on the subject including the evergreen Dantzig (1963), as well as the more recent Matousek and Gartner (2007) and Vanderbei (2010).

Linear programming is a subset of mathematical programming facing the efficient allocation of limited resources to known activities with the objective of meeting a desired goal, such as minimizing cost or maximizing profit.

The variables:

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

whose values are to be decided in some optimal fashion, are referred to as *decision variables.*

In a general linear program, the aim is to find a vector $\mathbf{x}^* \in \mathbb{R}^n_+$ minimizing (or maximizing) the value of a given linear function among all vectors $\mathbf{x} \in \mathbb{R}^n_+$ that satisfy a given system of linear equations and inequalities. The role of linearity is, therefore, twofold:

- the objective function, that is the quality of the plan, is measured by a linear function of the considered quantities;
- feasible plans are restricted by linear constraints (inequalities).

The linearity of some models can be justified on the basis of the typical properties of the problem. However, some nonlinear problems can be linearized by a proper use of mathematical transformations. This is, indeed, the case of the QR problem, as it will be shown in the next subsection. The representation (or sometimes the approximation) of a problem with a linear programming formulation ensures that efficient procedures are available for computing solutions.

In order to recap, a typical linear programming problem satisfies the following conditions:

- The n decision variables are non-negative:

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

Geometrically speaking, this restricts the solutions to $\mathbb{R}^n_+$.

In case the problem is characterized by variables unrestricted in sign, that is variables that can be positive, negative or zero, a simple trick can be used to restrict to non-negative variables, namely introducing two non-negative variables $[x]^+$ and $[-x]^+$ and converting any unconstrained decision variable as difference between two non-negative variables without changing the optimization problem:

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

For $x > 0$, we set $[x]^+ = x$ and $[-x]^+ = 0$, while for $x < 0$ we set $[-x]^+ = -x$ and $[x]^+ = 0$.

If $x = 0$, then both $[x]^+ = [-x]^+ = 0$.

- The criterion for choosing the optimal values of the decision variables, that is the objective function, is a linear function of the same variables:

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

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

- The m constraints regulating the process can be expressed as linear equations and/or linear inequalities written in terms of the decision variables. A generic constraint consists of either an equality or an inequality associated with some linear combinations of the decision variables:

$$a_1 x_1 + \cdots + a_i x_i + \cdots + a_n x_n \left\{ \begin{array}{c} \leq \\ = \\ \geq \end{array} \right\} b.$$

It is easy to convert constraints from one form to another. For example, an inequality constraint:

$$a_1 x_1 + \cdots + a_i x_i + \cdots + a_n x_n \leq b$$

can be converted to a greater than or equal constraint simply by multiplying it by -1. Instead, it can be converted to an equality constraint by adding a non-negative variable (*slack variable*):

$$a_1 x_1 + \cdots + a_i x_i + \cdots + a_n x_n + w = b, \quad w \geq 0.$$

On the other hand, an equality contraint:

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

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

$$a_1x_1 + \cdots + a_ix_i + \cdots + a_nx_n \leq b$$
$$a_1x_1 + \cdots + a_ix_i + \cdots + a_nx_n \geq b.$$

This is to say that there are no differences for how one poses the constraints.

Finally, it is worth recalling, that from a geometric point of view, a linear equation corresponds to a hyperplane, while 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.

By combining all the inequalities at once, the solution set (*feasible set*) is then the intersection of all the involved half-spaces: the n half-spaces posed by the non-negativity of the decision variables and the m half-spaces posed by the m constraints regulating the process.

Let us consider dealing with a linear programming problem with n unknowns (decision variables) and m constraints. The matrix formulation of the problem is the following (*standard form*):

$$\begin{aligned} &\text{minimize} && \mathbf{cx} \\ &\text{subject to} && \mathbf{Ax} \leq \mathbf{b} \\ & && \mathbf{x} \geq \mathbf{0}. \end{aligned} \tag{2.1}$$

The standard form poses the inequalities as a less than or equal form and requires the non-negativity for the decision variables.

The vector $\mathbf{c}_{[n]}$ contains the costs for the decision variables while the matrix $\mathbf{A}_{[m\times n]}$ and the vector $\mathbf{b}_{[m]}$ allow us to take into account the m constraints. A solution x is called feasible if it satisfies all the constraints. It is called optimal, and denoted by x^*, if it attains the minimum. The optimal vector $\mathbf{x}^*$ is therefore the feasible vector of least cost.

We have already stated how the condition $\mathbf{x} \geq \mathbf{0}$ restricts $\mathbf{x}$ to $\mathbb{R}^n_+$, that is to 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 half-spaces. The feasible set consists of the intersection of the above mentioned $m + n$ half-spaces. Such a feasible set can be bounded, unbounded or empty. The cost function $\mathbf{cx}$ produces a family of parallel planes, that is the plane $\mathbf{cx} = constant$ corresponds to the plane whose cost is equal to *constant*; when *constant* varies, the plane sweeps out the whole n-dimensional space. The optimal $\mathbf{x}^*$ is the $\mathbb{R}^n_+$ point, that is the n-dimensional vector, that ensures the lowest cost lying in the feasible set.

A problem that has no feasible solution is called infeasible. A problem with arbitrarily larger objective values is unbounded.

Associated with every linear program is its dual program. According to Matousek and Gartner (2007), duality theory is the most important theoretical result about linear programs. Starting from the linear problem introduced in Equation (2.1), also called the *primal problem*, its corresponding dual formulation starts with the same $\mathbf{A}$ and $\mathbf{b}$ but reverses anything else. The primal cost vector $\mathbf{c}$ and the constraint vector $\mathbf{b}$ are indeed switched to the dual profit vector $\mathbf{b}$ with $\mathbf{c}$ used as the constraint vector. The dual unknown (decision variable) $\mathbf{y}$ is now a vector with m components, the n constraints being represented by $\mathbf{yA} \geq \mathbf{c}$. Given, therefore, the linear programming problem (2.1), the associated *dual linear program* is:

$$\begin{aligned} &\text{maximize} && \mathbf{by} \\ &\text{subject to} && \mathbf{yA} \geq \mathbf{c} \\ & && \mathbf{y} \geq \mathbf{0}. \end{aligned} \tag{2.2}$$

Linear programs come then in primal/dual pairs. It turns out that every feasible solution for one of these two linear programs gives a bound on the optimal objective value for the other. Theoretical results (Matousek and Gartner 2007) ensure that:

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

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

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

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

Referring the reader to the references above for further details, it is worth noticing that it is sometimes easier to solve the linear program starting from the dual, rather than from the primal, formulation.

Finally, several authors express the inequality constraints in Equation (2.1) and Equation (2.2) as equalities (*equational form*), that is as $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{yA} = \mathbf{c}$, respectively, through the introduction of proper slack variables. The vectors $\mathbf{x}$ and $\mathbf{y}$ are then composed of the original decision variables and of the slack variables. Such formulation is used below for THE linear programming formulation of the QR problem.

2.1.2 The linear programming formulation for the QR problem

L_1 regression (Bloomfield and Steiger 1983), also known as median regression, is a natural extension of the sample median when the response is conditioned on the

covariates, as outlined in Section 1.1.3. The linear programming formulation for the conditional median is shown in the following, first for the simple regression model and then for the multiple regression model.

2.1.2.1 The two-variables problem

Let us first consider a two-variables problem. Searching the best line approximating the data scatter according to the L_1 criterion corresponds to the solution of the following minimization problem:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |\beta_0 + \beta_1 x_i - y_i|. \tag{2.3}$$

Although this cost function is not linear, a simple trick allows us to make it linear, at the price of introducing extra variables. In fact, an objective function or constraints involving absolute values can be handled via linear programming by introducing extra variables or extra constraints, as outlined above.

The resulting linear program is:

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

Each e_i is an auxiliary variable standing for the error at the i-th point. The constraints guarantee 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 Expression (2.3). Hence, to solve the L_1 regression problem, it suffices to solve the equivalent linear programming problem.

To illustrate the linear programming formulation of the L_1 problem, we refer again to the *Cars*93 dataset used for the samples in Chapter 1. For the sake of illustration, let us consider only the eight cars produced by the Ford manufacturer. The values for *Horsepower* (regressor) and *Price* (response variable) are reported in Table 2.1. The relationship between the two variables, along with the L_1 regression lines, are shown in Figure 2.1.

Table 2.1 *Horsepower* (regressor) and *Price* (response variable) for the eight cars produced by the Ford manufacturer (*Cars*93 dataset).

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

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

$$
\begin{array}{lrrrrrrrrrrl}
\text{minimize} & \sum_{i=1}^{10} e_i & & & & & & & & & & \\
\text{subject to} & - & \beta_0 - 63\,\beta_1 & -e_1 & & & & & & & & \leq -7.4 \\
& - & \beta_0 - 127\,\beta_1 & & -e_2 & & & & & & & \leq -10.1 \\
& - & \beta_0 - 96\,\beta_1 & & & -e_3 & & & & & & \leq -11.3 \\
& - & \beta_0 - 105\,\beta_1 & & & & -e_4 & & & & & \leq -15.9 \\
& - & \beta_0 - 115\,\beta_1 & & & & & -e_5 & & & & \leq -14.0 \\
& - & \beta_0 - 145\,\beta_1 & & & & & & -e_6 & & & \leq -19.90 \\
& - & \beta_0 - 140\,\beta_1 & & & & & & & -e_7 & & \leq -20.2 \\
& - & \beta_0 - 190\,\beta_1 & & & & & & & & -e_8 & \leq -20.9 \\
& & \beta_0 + 63\,\beta_1 & -e_1 & & & & & & & & \leq 7.4 \\
& & \beta_0 + 127\,\beta_1 & & -e_2 & & & & & & & \leq 10.1 \\
& & \beta_0 + 96\,\beta_1 & & & -e_3 & & & & & & \leq 11.3 \\
& & \beta_0 + 105\,\beta_1 & & & & -e_4 & & & & & \leq 15.9 \\
& & \beta_0 + 115\,\beta_1 & & & & & -e_5 & & & & \leq 14.0 \\
& & \beta_0 + 145\,\beta_1 & & & & & & -e_6 & & & \leq 19.9 \\
& & \beta_0 + 140\,\beta_1 & & & & & & & -e_7 & & \leq 20.2 \\
& & \beta_0 + 190\,\beta_1 & & & & & & & & -e_8 & \leq 20.9 \\
& & & e_1, & e_2, & e_3, & e_4, & e_5, & e_6, & e_7, & e_8 & \geq 0.
\end{array}
$$

The solution to this linear programming is:

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

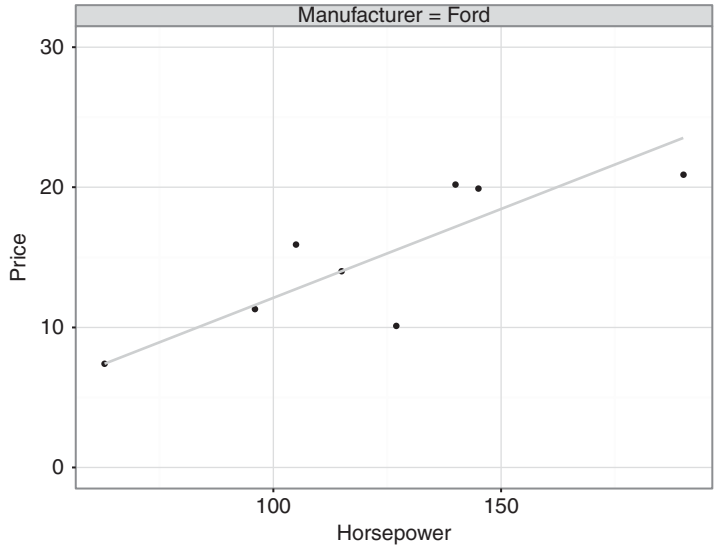

Figure 2.1 Relationship between Price *(response variable) and* Horsepower *(regressor) for the eight cars produced by the Ford manufacturer (*Cars*93 dataset). The line corresponds to the* L_1 *(median) regression line.*

2.1.2.2 The *p*-variables problem

The model for linear quantile regression is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}(\theta) + \boldsymbol{\epsilon},$$

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 and $\boldsymbol{\epsilon}_{[n]}$ is the vector of unknown errors. In the remainder of this section, the simpler notation $\boldsymbol{\beta}$ will be used to refer to the conditional median case ($\theta = 0.5$).

The least absolute estimates $\hat{\boldsymbol{\beta}}$ for the conditional median is obtained as the solution of the minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \boldsymbol{\beta}|. \tag{2.4}$$

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

$$\mathbf{s}_1 = [\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]_+$$
$$\mathbf{s}_2 = [\mathbf{X}\boldsymbol{\beta} - \mathbf{y}]_+$$

the original L_1 problem can be formulated as:

$$\min_{\boldsymbol{\beta}}\{\mathbf{1}^\top \mathbf{s}_1 + \mathbf{1}^\top \mathbf{s}_2 | \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{s}_1 - \mathbf{s}_2, \{\mathbf{s}_1, \mathbf{s}_2\} \in \mathbb{R}^n_+\}.$$

Furthermore, let:

$$\mathbf{B} = [\mathbf{X} - \mathbf{X}\mathbf{I} - \mathbf{I}],$$

and

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

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

Such reformulation of the problem leads to a standard linear programming problem. The primal formulation of such a problem (equational form) is:

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

Therefore, its dual counterpart is:

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

Bearing in mind the main result of linear programming, that is the theorem for which the solutions of such a minimization problem have to be searched in the vertices of the simplex[1], by a simple position, the above problem can be reformulated as follows:

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

[1] See the following subsection for more details.

In fact the equality:

$$\mathbf{X}^\top \mathbf{z} = \mathbf{0}$$

can be transformed as follows:

$$\tfrac{1}{2}\mathbf{X}^\top \mathbf{z} = \mathbf{0} \qquad \{\text{by multiplying by } \tfrac{1}{2}\}$$

$$\tfrac{1}{2}\mathbf{X}^\top \mathbf{z} + \tfrac{1}{2}\mathbf{X}^\top \mathbf{1} = \tfrac{1}{2}\mathbf{X}^\top \mathbf{1} \qquad \{\text{by adding } \tfrac{1}{2}\mathbf{X}^\top \mathbf{1}\}.$$

The obtained formulation:

$$\mathbf{X}^\top \underbrace{\left(\frac{1}{2}\mathbf{z} + \frac{1}{2}\mathbf{1}\right)}_{\boldsymbol{\eta}} = \underbrace{\frac{1}{2}\mathbf{X}^\top \mathbf{1}}_{\mathbf{b}} \tag{2.5}$$

permits the expression of the dual problem as follows:

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

The role of $1/2$ in Equation (2.5) is seemingly neutral, but it is the key to the generalization to the other conditional quantiles. In fact, the minimization problem for the conditional median, Equation (2.4), becomes for the generic θ-th conditional quantile:

$$\min_{\boldsymbol{\beta}(\theta)} \sum_{i=1}^{n} \rho_\theta (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\theta))$$

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

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

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

2.1.3 Methods for solving the linear programming problem

In 1947, Dantzig (1963) proposed the simplex method for solving linear programming problems. The simplex method is an iterative process, starting from a solution that satisfies the constraints and the non-negativities posed by the decision variables. It then 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.

An intuitive explanation of the simplex idea follows from the above introduced geometric explanation of the feasible set as the intersection of the $n + m$ half-spaces posed by the decision variables and the constraints. If the problem has a feasible solution, the optimal vector occurs at a corner of the feasible set, as guaranteed by the geometry of the problem: the planes corresponding to the cost function to be minimized in the primal formulation (the profit function to be maximized 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 movement along the edges of the feasible set: starting from a corner, the procedure goes from corner to corner of the feasible set until it finds the corner with the lowest (highest) associated cost (profit). At a typical corner, there are n edges to choose from, some leading away from the optimal solution, with others gradually moving toward it. Dantzig proposed choosing a corner corresponding to a lower cost (higher profit) and showed that there is no possibility to return to anything worse: the optimal solution, that is the corner $\mathbf{x}^*$, is reached when there is a corner from which all edges go the wrong way.

For the QR problem, the efficient version of the simplex algorithm proposed by Barrodale and Roberts (1974), adapted by Koenker and D'Orey (1987) to compute conditional quantiles, is typically used with a moderate size problem. It is, indeed, the default option in most of the QR software (SAS Institute Inc. 2010; Koenker 2011). It has been shown that the algorithm is slow with a large number of observations ($n > 100\ 000$) (Chen and Wei 2005).

A completely different method approaches the solution from the interior of the feasible set rather than on its boundary, that is starting in the zone of the set where all the inequalities are strictly satisfied (that is considering $<$ rather than $\leq$). Such methods, called interior-point methods, have their roots in the seminal paper of Karmakar (1984), that claimed a strong superiority for its proposal with respect to the simplex solution: he claimed that his version was 50 times faster than the simplex. Although his claim proved to be exaggerated, nonetheless interior-point methods have been shown to be competitive with the simplex. In particular, they are usually superior on very large problems.

Portnoy and Koenker (1997) proposed the use of interior-point methods for QR showing their efficiency in the case of datasets with a large number of observations.

Finally, it is worth mentioning a heuristic approach more recently proposed by Chen (2004, 2007). Such an algorithm, called the finite smoothing algorithm, is faster and more accurate for approximating the original problem with respect to the interior-point method in the presence of a large number of covariates. Typical options for QR software confirm this state of the art: they indeed automatically switch to (or suggest to use) the interior-point solution in the case of datasets with a large number of observations and the smoothing solution in the presence of a large number of covariates (SAS Institute Inc. 2010; Koenker 2011).

A very itemized presentation of all the technical details, as well as of the modern implementation techniques in order to take advantage of the different proposals, is given in the paper by Chen and Wei (2005). The paper covers both the estimation

issues and the inference issues, related to QR, from a computational point of view, offering details on the different algorithms and their performance.

2.2 A set of illustrative artificial data

In the following sections of this chapter, a set of artificial data is used to show several QR features. Such data share the same deterministic structure, but they differ with respect to the error term. The aim of their use is to show the QR behavior for different typologies of homogeneous and heterogeneous error terms and in the presence of dependence structures. The following subsections outline the different error terms and the corresponding models.

2.2.1 Homogeneous error models

Homogeneous regression models can be different with respect to the distribution of the error term, that is homogeneous error models include, but are not limited to, the classical normal regression model. In the latter, the error term follows a standardized normal distribution. Other models belonging to this class can be characterized by symmetric not normal errors, as well as by asymmetric errors. In order to show some of them, we consider the following random variables:

- a standard normal variable:

$$\mathbf{e}_N \sim N(\mu = 0, \sigma = 1);$$

- three log-normal variables with location parameter $\mu = 0$, but with increasing scale parameter σ:

$$\mathbf{e}_{LN_1} \sim LN(\mu = 0, \sigma = 0.25);$$

$$\mathbf{e}_{LN_2} \sim LN(\mu = 0, \sigma = 0.5);$$

$$\mathbf{e}_{LN_3} \sim LN(\mu = 0, \sigma = 1.25).$$

The densities for the random variables, $\mathbf{e}_N$, $\mathbf{e}_{LN_1}$, $\mathbf{e}_{LN_2}$ and $\mathbf{e}_{LN_3}$, are shown in Figure 2.2.

Starting from such errors, we define the following four models:

- a homogeneous error model with a normal error term:

$$model_1 \rightarrow \mathbf{y}^{(1)} = 1 + 2\mathbf{x} + \mathbf{e}_N;$$

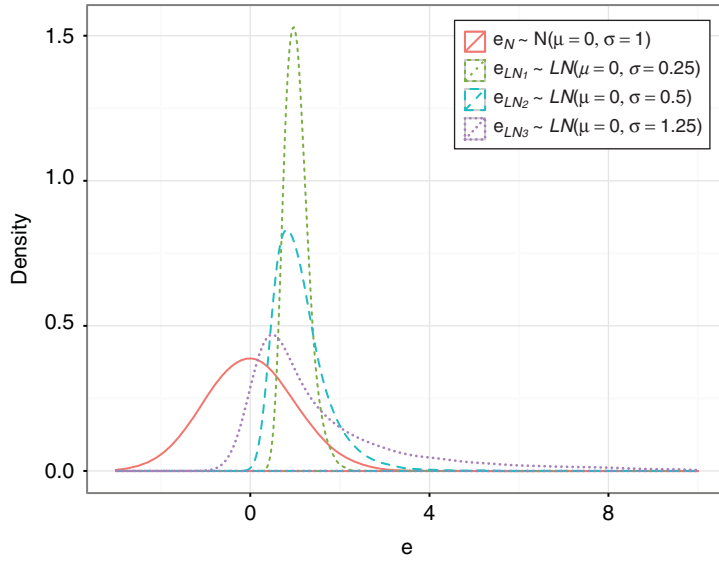

Figure 2.2 Three examples of error distributions (dashed lines) that can characterize a homogeneous error model, along with the standard normal curve (solid line). The three curves refer to error terms simulated from a log-normal distribution with location parameter $\mu = 0$ and scale parameter $\sigma = 0.25$ (the not normal symmetric density) and from a log-normal distribution with location parameter $\mu = 0$ and increasing scale parameter $\sigma = \{0.5, 1.25\}$ (the asymmetric densities).

- a homogeneous error model with a symmetric but not normal error term:

$$model_2 \rightarrow \mathbf{y}^{(2)} = 1 + 2\mathbf{x} + \mathbf{e}_{LN_1};$$

- three homogeneous error models characterized by asymmetric error terms with an increasing skewness ($model_3$ and $model_4$) and by a different effect ($model_5$):

$$model_3 \rightarrow \mathbf{y}^{(3)} = 1 + 2\mathbf{x} + \mathbf{e}_{LN_2};$$
$$model_4 \rightarrow \mathbf{y}^{(4)} = 1 + 2\mathbf{x} + \mathbf{e}_{LN_3};$$
$$model_5 \rightarrow \mathbf{y}^{(5)} = 1 + 2\mathbf{x} - \mathbf{e}_{LN_3}.$$

While $model_1$ respects the assumptions of the classical normal regression framework (Gujarati 2003), the other four models are characterized by a violation of the classical assumptions with respect to the presence of a symmetric but not normal error ($model_2$), or asymmetric errors with different degrees of skewness ($model_3$ and $model_4$). Finally, $model_5$ has the same error term as $model_4$, but it enters the model with a different sign. A random sample of $n = 10\,000$ observations extracted from each model is represented by scatterplots in Figure 2.3.

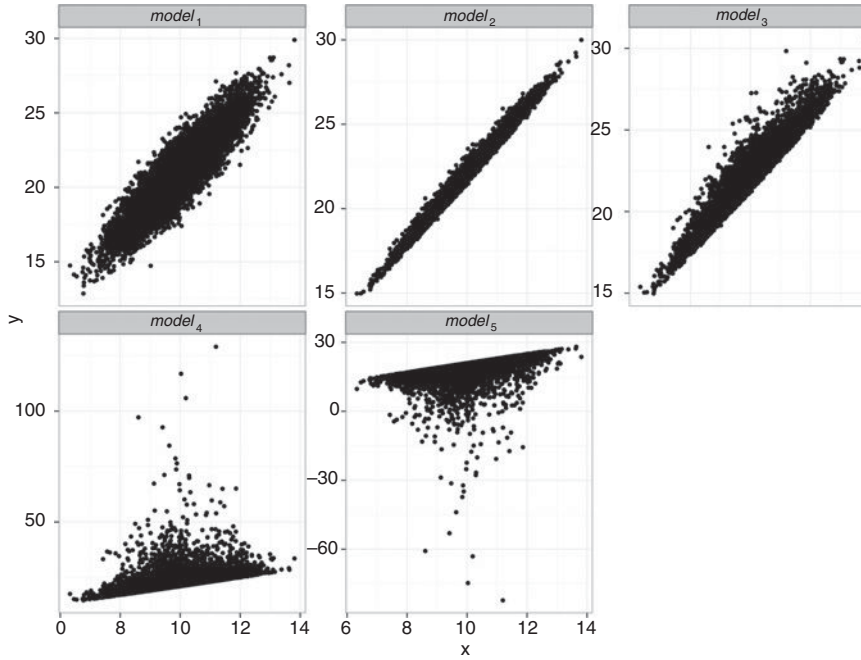

Figure 2.3 Scatterplots for the five models, model_1, model_2, model_3, model_4 *and* model_5, *clockwise from the top-left corner.* model_1 *and* model_2 *show homoskedastic patterns, although only* model_1 *is characterized by a normal error, while the error term of* model_2 *has a symmetric not normal shape.* model_3 *and* model_4 *are, instead, characterized by an increasing skewness. Finally,* model_5 *has the same error term as* model_4, *but here it is considered with the opposite sign. For each model, a random sample of* $n = 10\,000$ *observations is represented.*

2.2.2 Heterogeneous error models

A violation of the homoskedastic assumption is probably the most prevalent point for the use of QR, although, as we will show, it is not the only case in which such a technique is useful. In order to take into account a simple heteroskedastic pattern, we consider the following model, starting from the standard normal error term:

$$model_6 \rightarrow \mathbf{y}^{(6)} = 1 + 2\mathbf{x} + (1 + \mathbf{x})\mathbf{e}_N.$$

Figure 2.4 depicts the scatterplot for a random sample of $n = 10\,000$ units extracted from $model_6$. From the analysis of the plot, the increasing variance of the error term with the values of the regressor is evident. It is naturally possible to take into account different degrees and shapes of heteroskedastic error, but the use of such a model is sufficient for the illustrative aim of this chapter.

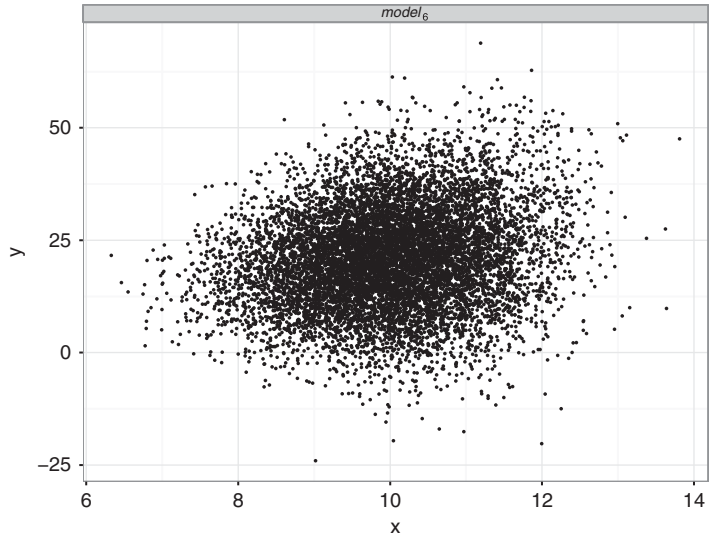

Figure 2.4 Scatterplot for a random sample extracted from the heterogeneous model (model_6*): the error term increases with the regressor values.*

2.2.3 Dependent data error models

Four further models are used to show the QR behavior with respect to the error independence assumption. In particular, starting from an autoregressive error term of order 1:

$$\mathbf{e}_{AR(1)_{[rho]}} \rightarrow e_i = \rho e_{i-1} + a_i$$

where $a_i \sim N(\mu = 0, \sigma = 1)$ and using the values $\rho = \{-0.2, +0.2, -0.5, +0.5\}$, we define the following four models:

$$model_7 \rightarrow \mathbf{y}^{(7)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=-0.2]};$$
$$model_8 \rightarrow \mathbf{y}^{(8)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=+0.2]};$$
$$model_9 \rightarrow \mathbf{y}^{(9)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=-0.5]};$$
$$model_{10} \rightarrow \mathbf{y}^{(10)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=+0.5]}.$$

The four error densities for the different values of ρ are shown in Figure 2.5: the two higher curves at the median correspond to $\rho = \pm 0.2$, while the curves for $\rho = \pm 0.5$ are characterized by heavier tails.

The 10 models introduced above will be used throughout the chapter in order to show some of the main QR features. An outline summary of the different models is offered in Table 2.2, where the different models are classified according to the type of model (first column), the error term (second column) and the data generating process (third column). For all the models $\mathbf{x} \sim N(\mu = 10, \sigma = 1)$ is used.

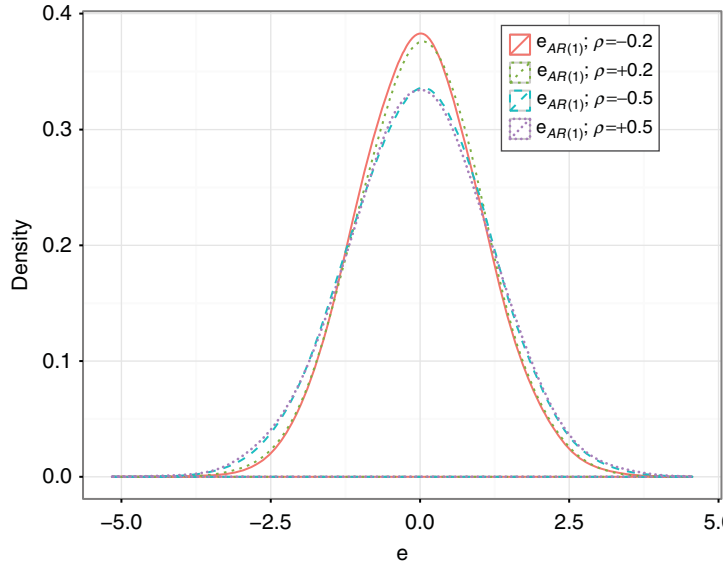

Figure 2.5 Four examples of dependent error distributions generated using an AR(1) model with different values of the autocorrelation coefficient.

Table 2.2 An outline summary for the different illustrative models used in this chapter: the first column shows the type of error model, the second column shows the error term, while the resultant model is in the third column.

Error model	Error term	$model_i$
Homogeneous	$\mathbf{e}_N \sim N$ $(\mu = 0, \sigma = 1)$	$model_1 \rightarrow \mathbf{y}^{(1)} = 1 + 2\mathbf{x} + \mathbf{e}_N$
	$\mathbf{e}_{LN_1} \sim LN$ $(\mu = 0, \sigma = 0.25)$	$model_2 \rightarrow \mathbf{y}^{(2)} = 1 + 2\mathbf{x} + \mathbf{e}_{(LN_1)}$
	$\mathbf{e}_{LN_2} \sim LN$ $(\mu = 0, \sigma = 0.5)$	$model_3 \rightarrow \mathbf{y}^{(3)} = 1 + 2\mathbf{x} + \mathbf{e}_{(LN_2)}$
	$\mathbf{e}_{LN_3} \sim LN$ $(\mu = 0, \sigma = 1.25)$	$model_4 \rightarrow \mathbf{y}^{(4)} = 1 + 2\mathbf{x} + \mathbf{e}_{(LN_3)}$
		$model_5 \rightarrow \mathbf{y}^{(5)} = 1 + 2\mathbf{x} - \mathbf{e}_{(LN_3)}$
Heterogeneous	$\mathbf{e}_N \sim N$ $(\mu = 0, \sigma = 1)$	$model_6 \rightarrow \mathbf{y}^{(6)} = 1 + 2\mathbf{x} + (1 + \mathbf{x})\mathbf{e}_N$
Dependent	$\mathbf{e}_{AR(1)[\rho=-0.2]}$	$model_7 \rightarrow \mathbf{y}^{(7)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=-0.2]}$
	$\mathbf{e}_{AR(1)[\rho=+0.2]}$	$model_8 \rightarrow \mathbf{y}^{(8)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=+0.2]}$
	$\mathbf{e}_{AR(1)[\rho=-0.5]}$	$model_9 \rightarrow \mathbf{y}^{(9)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)[\rho=-0.5]}$
	$\mathbf{e}_{AR(1)[\rho=+0.5]}$	$model_{10} \rightarrow \mathbf{y}^{(10)} = 1 + 2\mathbf{x} + \mathbf{e}_{AR(1)_{[\rho=+0.5]}}$

2.3 How and why to work with QR

2.3.1 QR for homogeneous and heterogeneous models

This section focuses on the interpretation of the QR estimated coefficients by drawing a parallel between homogeneous and heterogeneous regression models. Starting from a comparison of QR behavior on such different models, its added value will be highlighted. In fact, while it is easy to understand how QR is important in dealing with heteroskedasticity, the use of different distributions for the error term will offer the opportunity to observe QR capability to estimate the whole conditional distribution also in the presence of homoskedasticity.

In particular, in order to appreciate the added value of QR, it is useful to compare the QR estimates in the case of homogeneous variance regression models, and in the case of heterogeneous variance regression models. For this purpose, let us consider $model_1$ and $model_6$ introduced above: the two corresponding scatterplots, along with the conditional mean fit, conditional median fit, and conditional quantile fits, are represented for both models in Figure 2.6, for $\theta = \{0.05, 0.1, 0.25, 0.75, 0.9, 0.95\}$. For the homogeneous variance regression $model_1$ [Figure 2.6(a)], the only estimated effect is a change in the central tendency of the distribution of Y conditional on the value of $\mathbf{x}$ (location model). QR slope estimates are then the same at all QRs, and any deviation among the regression estimates is simply due to sampling variation: an estimate of the rate of change in the mean from ordinary least squares (OLS) regression is also an estimate of the same parameter as for the QR. When the predictor variable $\mathbf{x}$ exerts both a change in mean and a change in variance on the distribution of Y (location-scale model), changes in the quantiles of Y cannot be the same for all the quantiles, as for the case of $model_6$. Figure 2.6(b) shows that slope estimates differ across quantiles since for this model the variance in Y changes as a function of $\mathbf{x}$. Thus, in such a case, OLS regression analysis provides an incomplete picture of the relationship between variables, as it only focuses on changes at the conditional mean.

The above discussion shows the importance of QR in order to describe the entire conditional distribution of a dependent variable. In the heterogeneous setting, QR coefficients differ in magnitude (or size) and direction (or sign), thus providing location, scale and shape shift information on Y. Using the whole quantile process (see Section 2.3.2), indeed, the conditional densities at different values of X can be recovered: this allows to investigate both the scale and the shape effect. The reader is referred to Chapter 4, Section 4.2 for details on the conditional density estimation. It is, therefore, obvious to claim that QR offers a more complete view of relationships among variables for heterogeneous regression models, providing a method for modeling the rates of changes in the response variable at multiple points of the distribution when such rates of change are different. QR is, however, also a useful tool in the case of homogeneous regression models outside of the classical normal regression model. When the error term satisfies the classical normal assumptions (Gujarati 2003), the QR estimates can be obtained simply by adding/substracting the corresponding standardized normal quantiles from the OLS slope estimate, as the slope estimate does not vary across the quantiles of the conditional distribution. By referring to the homogeneous error $model_1$ introduced above, Figure 2.7

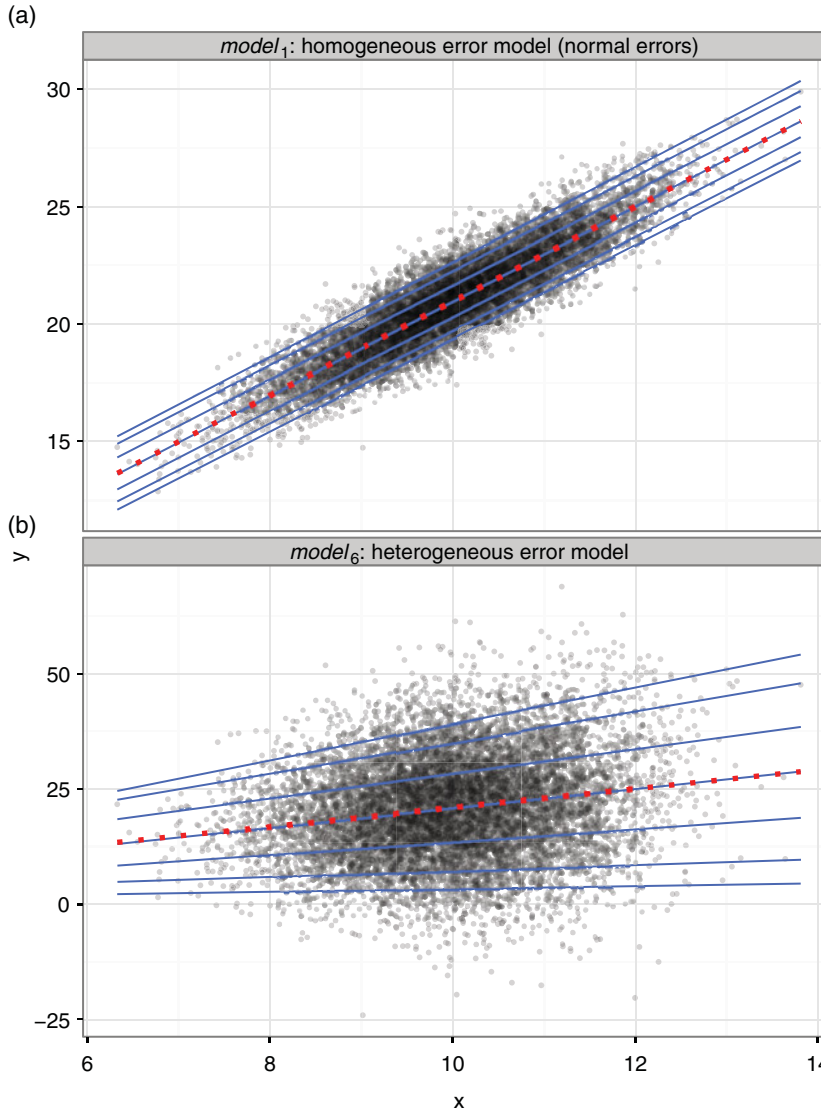

Figure 2.6 Scatterplots, OLS line (dotted line), and conditional quantile lines (solid lines), $\theta = \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$, for a homogeneous error model (a) and for a heterogeneous error model (b). For the homogeneous error model_1 *(location model), the slope estimate is constant across conditional quantiles, while the differences are in the intercept estimates. In the heterogeneous error* model_6 *(location-scale model), the predictor variable X exerts both a change in mean and a change in variance on the distribution of Y, thus causing differences across conditional quantiles both in intercept estimates and in slope estimates.*

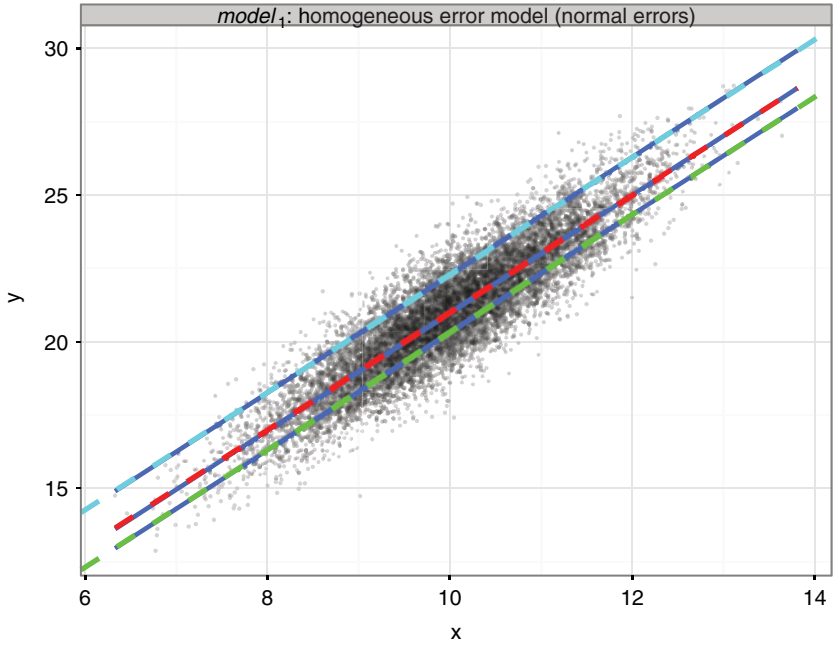

Figure 2.7 Scatterplot, OLS line (middle dashed line) and conditional quantile lines for $\theta = \{0.25, 0.5, 0.9\}$ *(solid lines). In the case of a homogeneous error model with normal errors, the QR estimates can be easily obtained by shifting the OLS line using the corresponding quantiles of the standard normal distribution. The two external dashed lines have been obtained simply by shifting the OLS line by* -0.67 *(first quartile of a standard normal distribution) and by +1.28 (90-th percentile of a standard normal distribution), respectively.*

superimposes on the data scatterplot the QR estimates for $\theta = \{0.25, 0.5, 0.9\}$ (solid lines), and the lines obtained by shifting the OLS line (middle dashed line) by -0.67 (first quartile of a standard normal distribution) and by +1.28 (90-th percentile of a standard normal distribution). The two external dashed lines coincide with the QR lines, thus making unnecessary the whole QR machinery in such a setting.

On the contrary, the situation is different for different types of homogeneous models, where the normal distribution does not hold. Figure 2.8 depicts the two scatterplots for $model_2$ and $model_3$: from the comparison of the QR lines for $\theta = \{0.25, 0.9\}$ (external solid lines) with the ones obtained by shifting the OLS line using the corresponding quantiles of the standard normal curve (external dotted lines), strong differences are evident. Only a change in the error distribution assumption could lead to a correct estimation of the Y conditional distribution. QR is, therefore, a valid tool for estimating the entire conditional distribution of the dependent variable without requiring a prior analysis of the distribution of errors; that is through the adoption of a goodness of fit test.

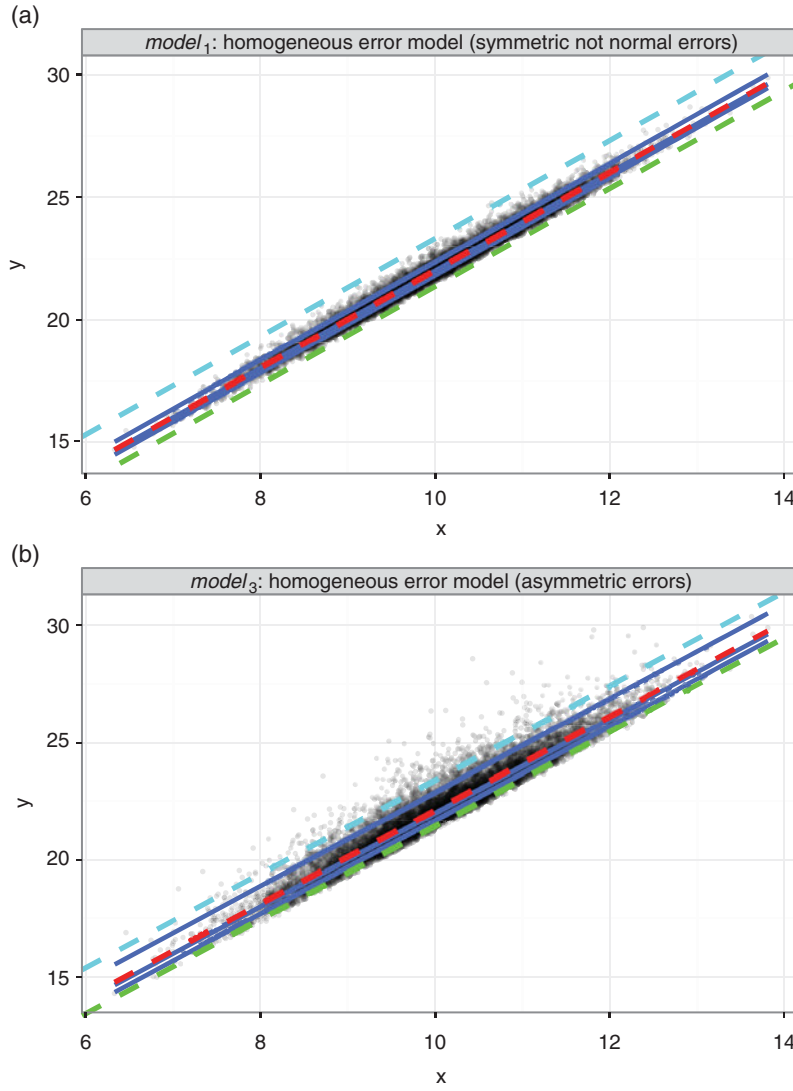

Figure 2.8 Scatterplot, OLS line (middle dashed line), and conditional quantile lines for $\theta = \{0.25, 0.5, 0.9\}$ *(solid lines) for two not normal homogeneous regression models: a model obtained using a symmetric not normal error term (a) and a model characterized by an asymmetric error term (b). In such cases, the shift of the OLS lines using the quantiles of the standard normal distribution (moving under the classical normal assumption for the error term) leads to mis-estimate of the real conditional quantiles of the response variable. The two external dashed lines refer to the first conditional quartile and to the 90-th conditional percentile under the classical normal assumption (see Figure 2.7 for the case of a homogeneous error model with normal error term).*

Finally, it is worth noticing that no parametric distribution assumption is required for the error: the study of QR lines and the comparison with the lines obtained using the classical normal assumption can then also provide useful hints on the nature of the error term. Such nonparametric feature, along with the least absolute criterion on which the linear QR estimator is based upon, ensures its robustness with respect to skewed tails and departures from normality. However, it is worth highlighting that the QR estimator can be very sensitive to outliers in the explanatory variable (He *et al.* 1990). Several proposals in the literature attempt to obtain a more robust form of QR: see, among others, Rousseeuw and Hubert (1999), Giloni *et al.* (2006) and, more recently, Neykov *et al.* (2012).

Further insights on the gain offered by the nonparametric QR character are presented in the following section.

2.3.2 QR prediction intervals

QR estimates can also be used to obtain prediction intervals for the conditional distribution of the response variable. This approach exploits the nonparametric nature of QR, and offers a flexible tool for estimating prediction intervals at any specified value of the explicative variables. Unlike the ordinary prediction intervals used in OLS regression (Gujarati 2003; Weisberg 2005), such intervals are not sensitive to departures from the distributional assumptions of the classical normal linear model.

For a given model, the interval provided by two distinct quantile estimates, $\hat{q}_Y(\theta_1, X = x)$ and $\hat{q}_Y(\theta_2, X = x)$, at any specified value of the regressor X, is a $(\theta_2 - \theta_1)\%$ prediction interval for a single future observation. If we consider, for example, $\theta_1 = 0.1$ and $\theta_2 = 0.9$, then the interval $\left[\hat{q}_Y(\theta_1 = 0.1, X = x); \hat{q}_Y(\theta_2 = 0.9, X = x)\right]$ can be interpreted as an 80% prediction interval corresponding to the given value $X = x$. The use of such a prediction interval ensures a coverage level consistent with the nominal level in spite of the nature of the error term, as QR does not suffer departures from the classical assumptions of the normal linear model. Moreover, such a coverage level is stable with respect to the specific value chosen for x. This is a great advantage with respect to classical regression, where the correspondence between the empirical and the nominal coverage levels is ensured only in the case of the usual distributional assumptions ($model_1$), while it downgrades when the degree of violation of these assumptions increases.

A comparison between nominal and empirical coverage levels for the 10 models introduced above permits to appreciate the use of the QR estimates at this aim. Table 2.3 reports the population intervals for the 10 models using a confidence level $(1 - \alpha) = 80\%$. Each row refers to a different $Y|X = x$ conditional distribution, that is to the 10 models, while the columns report five distinct values of X, (8, 9, 10, 11, 12), spanning along the range of the regressor variable. If we consider, for example, the interval for $model_1$ when $X = 8$, we obtain:

$$[q_Y(\theta = 0.1, X = 8); q_Y(\theta = 0.9, X = 8)]_{model_1} = [15.72; 18.28].$$

Table 2.3 10-th percentile and 90-th percentile, $[q_Y(\theta = 0.1, X = x); q_Y(\theta = 0.9, X = x)]$, of the population conditional distribution $Y|X = x$ in correspondence with five distinct values of X (columns) spanning the range of the regressor for the 10 models (rows) characterized by different error terms. The percentiles for the dependent error models (from $model_7$ to $model_{10}$) have been computed through simulations using 1 000 000 replications of random samples with $n = 1000$ from each model.

	$[q_Y(\theta = 0.1, X = x); q_Y(\theta = 0.9, X = x)]$				
	$x = 8$	$x = 9$	$x = 10$	$x = 11$	$x = 12$
$model_1$	[15.72; 18.28]	[17.72; 20.28]	[19.72; 22.28]	[21.72; 24.28]	[23.72; 26.28]
$model_2$	[17.73; 18.38]	[19.73; 20.38]	[21.73; 22.38]	[23.73; 24.38]	[25.73; 26.38]
$model_3$	[17.53; 18.90]	[19.53; 20.90]	[21.53; 22.90]	[23.53; 24.90]	[25.53; 26.90]
$model_4$	[17.20; 21.96]	[19.20; 23.96]	[21.20; 25.96]	[23.20; 27.96]	[25.20; 29.96]
$model_5$	[12.04; 16.80]	[14.04; 18.80]	[16.04; 20.80]	[18.04; 22.80]	[20.04; 24.80]
$model_6$	[5.47; 28.53]	[6.18; 31.82]	[6.90; 35.10]	[7.62; 38.38]	[8.34; 41.66]
$model_7$	[15.69; 18.31]	[17.69; 20.31]	[19.69; 22.31]	[21.69; 24.31]	[23.69; 26.31]
$model_8$	[15.69; 18.30]	[17.69; 20.30]	[19.69; 22.30]	[21.69; 24.30]	[23.69; 26.30]
$model_9$	[15.52; 18.48]	[17.52; 20.48]	[19.52; 22.48]	[21.52; 24.48]	[23.52; 26.48]
$model_{10}$	[15.52; 18.48]	[17.52; 20.48]	[19.52; 22.48]	[21.52; 24.48]	[23.52; 26.48]

Thus, for $model_1$, we have that 80% of the conditional distribution of $Y|X = 8$ is inside the interval [15.72; 18.28]. The percentiles for the dependent error models, that is for $model_7$, $model_8$, $model_9$ and $model_{10}$, are obtained via simulations. To this end, 1 000 000 replications of random samples with $n = 1000$ have been used.

Using the same level, $(1 - \alpha) = 80\%$, Table 2.4 shows the prediction intervals for a random sample extracted from each of the 10 models: the rows of the table report the OLS and QR results, while the columns refer to five distinct values of the regressor; each cell of the table illustrates the prediction interval and the corresponding width, that is the distance between the two extremes of the interval. From the analysis of the table, the QR method provides intervals of width less than or similar to the OLS intervals for all the samples; the biggest differences are for $model_4$ and $model_5$, that is in the case of a strong skew error term.

The results shown in Table 2.5 are of greater interest for the purposes of comparison between the two methods:

- 1000 random samples were generated starting from each model;
- for each sample the prediction intervals were computed using both the OLS and the QR method;
- each cell of the table shows the percentage of times the simulated prediction intervals cover the real population intervals.

The obtained percentages show how QR prediction intervals offer an empirical coverage level consistent with the nominal one for all the models, in spite of the nature of the error term. The rows for OLS prediction intervals indicate their underperformance in the case of a violation of the normal classical framework. The same results are graphically shown in Figure 2.9: the left panel shows the OLS empirical coverage rates for the 10 models (different lines), while the QR empirical coverages are depicted in the right panel. The horizontal line at $y = 0.8$ denotes the nominal coverage level. The graph immediately illustrates the best performance of QR prediction intervals: the empirical coverage lines are very close to the nominal levels for almost all 10 models in the QR panel, minor differences are present in the case of dependent error models.

In order to better investigate the differences between OLS and QR prediction intervals, the same results are shown in Figure 2.10, this time arranging the 10 models on the panels: in each panel, the solid line depicts the OLS empirical coverage level, while the QR level is shown using a dashed line. The horizontal line at $y = 0.8$ again denotes the nominal coverage level. This representation allow us to better appreciate how the OLS empirical level is close to the nominal level only for $model_1$. Moving from $model_2$ to $model_5$, the OLS empirical coverage level downgrades with the increasing skewness of the error terms, while QR prediction intervals continue to ensure a good match between the nominal and the empirical levels. The pattern for $model_6$ deserves a separate discussion, where the OLS coverage level is very low for the smaller X values, while grows considerably in correspondence to the

Table 2.4 Prediction intervals for the 10 illustrative models (rows of the table) at five distinct values of X (columns of the table). For each sample the OLS and the QR prediction intervals at the level $(1-\alpha) = 80\%$ are shown along with their corresponding width. The five distinct values of X have been chosen to cover the whole range of the regressor.

		$[\hat{q}_Y(\theta = 0.1, X = x); \hat{q}_Y(\theta = 0.9, X = x)]$ **Width**				
		$x = 8$	$x = 9$	$x = 10$	$x = 11$	$x = 12$
$model_1$	LS	[15.82 ; 18.42] **2.60**	[17.79 ; 20.36] **2.57**	[19.75 ; 22.31] **2.56**	[21.69 ; 24.27] **2.57**	[23.63 ; 26.24] **2.60**
	QR	[15.54; 18.23] **2.69**	[17.78 ; 20.22] **2.53**	[19.83 ; 22.20] **2.38**	[21.97 ; 24.19] **2.22**	[24.11 ; 26.17] **2.06**
$model_2$	LS	[17.73 ; 18.41] **0.67**	[19.72 ; 20.39] **0.67**	[21.71 ; 22.37] **0.66**	[23.69 ; 24.35] **0.67**	[25.67 ; 26.34] **0.67**
	QR	[17.69 ; 18.36] **0.67**	[19.72 ; 20.36] **0.64**	[21.75 ; 22.35] **0.60**	[23.77 ; 24.35] **0.57**	[25.80 ; 26.34] **0.54**
$model_3$	LS	[17.48 ; 18.97] **1.49**	[19.45 ; 20.92] **1.47**	[21.41 ; 22.88] **1.47**	[23.36 ; 24.83] **1.47**	[25.31 ; 26.80] **1.49**
	QR	[17.48 ; 18.85] **1.37**	[19.52 ; 20.84] **1.32**	[21.56 ; 22.82] **1.27**	[23.60 ; 24.81] **1.21**	[25.64 ; 26.80] **1.16**
$model_4$	LS	[15.95 ; 23.18] **7.24**	[17.74 ; 24.89] **7.16**	[19.50 ; 26.63] **7.13**	[21.23 ; 28.39] **7.16**	[22.94 ; 30.18] **7.24**
	QR	[17.15 ; 21.65] **4.49**	[19.19 ; 23.57] **4.38**	[21.23 ; 25.49] **4.26**	[23.28 ; 27.42] **4.14**	[25.32 ; 29.34] **4.02**
$model_5$	LS	[10.82 ; 18.05] **7.24**	[13.11 ; 20.26] **7.16**	[15.37 ; 22.50] **7.13**	[17.61 ; 24.77] **7.16**	[19.82 ; 27.06] **7.24**

(continued)

Table 2.4 (cont'd)

		$[\hat{q}_Y(\theta = 0.1, X = x); \hat{q}_Y(\theta = 0.9, X = x)]$ **Width**				
		$x = 8$	$x = 9$	$x = 10$	$x = 11$	$x = 12$
$model_6$	QR	[12.35 ; 16.85] **4.49**	[14.43 ; 18.81] **4.38**	[16.51 ; 20.77] **4.26**	[18.58 ; 22.72] **4.14**	[20.66 ; 24.68] **4.02**
	LS	[3.97 ; 32.14] **28.17**	[5.71 ; 33.59] **27.87**	[7.36 ; 35.13] **27.78**	[8.90 ; 36.78] **27.89**	[10.33 ; 38.53] **28.20**
$model_7$	QR	[3.48 ; 28.13] **24.65**	[5.86 ; 31.17] **25.30**	[8.25 ; 34.20] **25.95**	[10.64 ; 37.24] **26.59**	[13.03 ; 40.27] **27.24**
	LS	[15.64 ; 18.44] **2.79**	[17.66 ; 20.42] **2.76**	[19.66 ; 22.42] **2.75**	[21.66 ; 24.42] **2.77**	[23.64 ; 26.44] **2.80**
$model_8$	QR	[16.31 ; 17.99] **1.68**	[18.12 ; 20.06] **1.94**	[19.93 ; 22.13] **2.20**	[21.74 ; 24.20] **2.46**	[23.55 ; 26.26] **2.72**
	LS	[15.67 ; 18.40] **2.73**	[17.70 ; 20.40] **2.70**	[19.72 ; 22.41] **2.69**	[21.72 ; 24.42] **2.70**	[23.72 ; 26.45] **2.73**
$model_9$	QR	[16.09 ; 18.08] **1.99**	[17.91 ; 20.19] **2.28**	[19.73 ; 22.29] **2.56**	[21.55 ; 24.40] **2.85**	[23.37 ; 26.50] **3.13**
	LS	[15.84 ; 18.90] **3.05**	[17.68 ; 20.70] **3.02**	[19.50 ; 22.51] **3.01**	[21.31 ; 24.33] **3.02**	[23.11 ; 26.16] **3.06**
$model_{10}$	QR	[16.20 ; 18.49] **2.29**	[17.82 ; 20.37] **2.55**	[19.43 ; 22.25] **2.81**	[21.05 ; 24.12] **3.07**	[22.66 ; 26.00] **3.33**
	LS	[15.82 ; 18.65] **2.83**	[17.73 ; 20.53] **2.80**	[19.63 ; 22.42] **2.79**	[21.52 ; 24.32] **2.80**	[23.40 ; 26.23] **2.83**
	QR	[15.87 ; 18.40] **2.53**	[17.76 ; 20.36] **2.60**	[19.65 ; 22.32] **2.67**	[21.54 ; 24.28] **2.74**	[23.43 ; 26.24] **2.81**

Table 2.5 Empirical coverage levels for OLS and QR prediction intervals computed using 1000 random samples extracted from each of the 10 considered models (rows of the table). The intervals are computed for five distinct values of X (columns of the table) to cover the whole range of the regressor.

		Empirical coverage level [Nominal coverage level $(1-\alpha) = 80\%$]				
		$x = 8$	$x = 9$	$x = 10$	$x = 11$	$x = 12$
$model_1$	LS	81.6	78.7	76.4	78.7	81.4
	QR	80.3	82.0	81.6	78.6	78.1
$model_2$	LS	75.3	69.2	65.5	70.1	74.3
	QR	80.4	82.0	81.4	79.0	77.9
$model_3$	LS	56.2	53.4	53.2	54.1	57.4
	QR	80.2	82.5	81.2	79.3	78.0
$model_4$	LS	18.2	17.8	17.1	17.1	17.9
	QR	80.6	82.1	81.2	79.2	78.4
$model_5$	LS	18.2	17.8	17.1	17.1	17.9
	QR	80.6	82.1	81.2	79.2	78.4
$model_6$	LS	26.5	38.8	74.6	95.3	99.4
	QR	80.0	81.4	81.4	79.4	78.1
$model_7$	LS	80.1	75.3	68.9	76.2	79.0
	QR	78.5	77.6	75.9	79.2	78.9
$model_8$	LS	84.0	81.0	77.5	82.1	82.4
	QR	80.8	81.8	81.9	81.1	79.9
$model_9$	LS	75.9	70.2	61.2	71.3	76.0
	QR	76.9	75.3	72.1	76.0	78.8
$model_{10}$	LS	84.3	83.6	82.9	83.3	84.1
	QR	80.6	83.0	85.5	82.5	81.5

highest values. Also in this case the QR coverage level is consistent with the nominal one. Finally, the two methods provide closer results for the error dependent models, that is from $model_7$ up to $model_{10}$, although also for these models, QR outperforms OLS. It is interesting to note how the coverage levels are lower with respect to the nominal one in the case of the two models with negative autocorrelation coefficients ($model_7$ and $model_9$), the opposite in the case of models with positive ρ ($model_8$ and $model_{10}$).

At last, in order to have a full comparison between the two methods, the average interval width has been computed on the same random samples used for the results shown in Table 2.5 and in Figures 2.9 and 2.10. The average interval widths, shown in Table 2.6, indicate how QR intervals offer better results also with respect to the interval precision, providing narrower, and therefore, more informative, intervals. Also, for the interval width, the strong differences between the OLS and QR methods

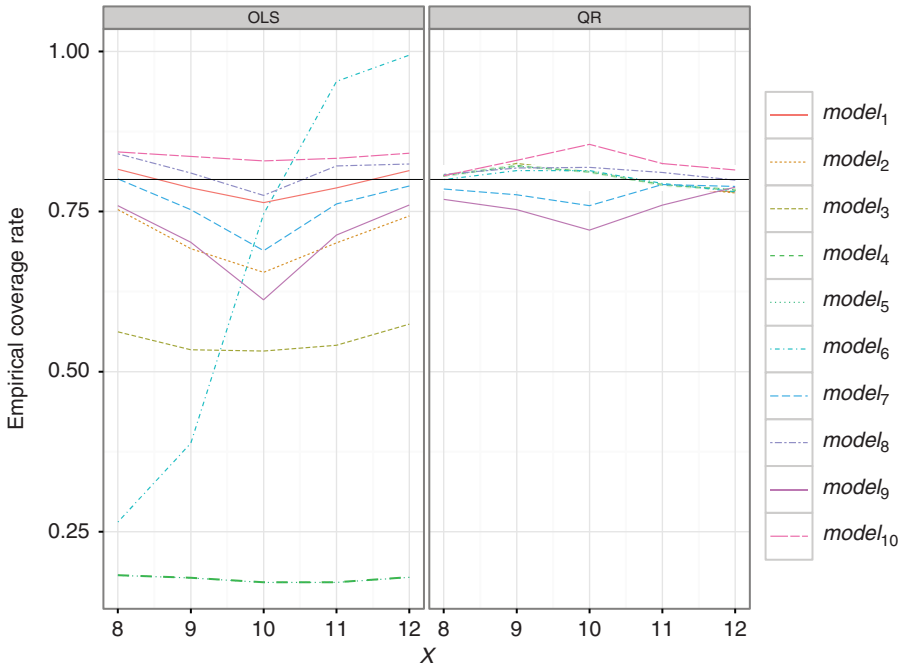

Figure 2.9 Empirical coverage levels computed on 1000 simulated datasets from the 10 different models, represented with different lines, using the classical OLS prediction intervals at the level $(1 - \alpha) = 80\%$ *(left panel) and two regression quantile estimates* $[\hat{q}_Y(\theta = 0.1, X = x); \hat{q}_Y(\theta = 0.9, X = x)]$ *(right panel) at five distinct values of* X *(abscissa). The horizontal line at* $y = 0.8$ *denotes the nominal coverage level.*

in the case of the models with skew error terms ($model_3$, $model_4$ and $model_5$) and for the heterogeneous error model ($model_6$) are very evident.

For a substantial treatment of various intervals based on regression quantiles, the interested reader may refer also to Zhou and Portnoy (1996) and to Koenker (2005).

2.3.3 A note on the quantile process

In the examples shown in Chapter 1 and in most applications discussed in this book, QR solutions are computed for a selected number of quantiles, typically the three quartiles along with two extreme quantiles, that is for $\theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. This is in light of the search for a rightful compromise between the amount of output to manage and the results to interpret and summarize. Although in many practical applications of QR, the focus is on estimating a subset of quantiles, it is worth noticing that it is possible to obtain estimates across the entire interval of conditional quantiles. In particular, the set:

$$\{\beta(\theta) : \theta \in (0, 1)\}$$

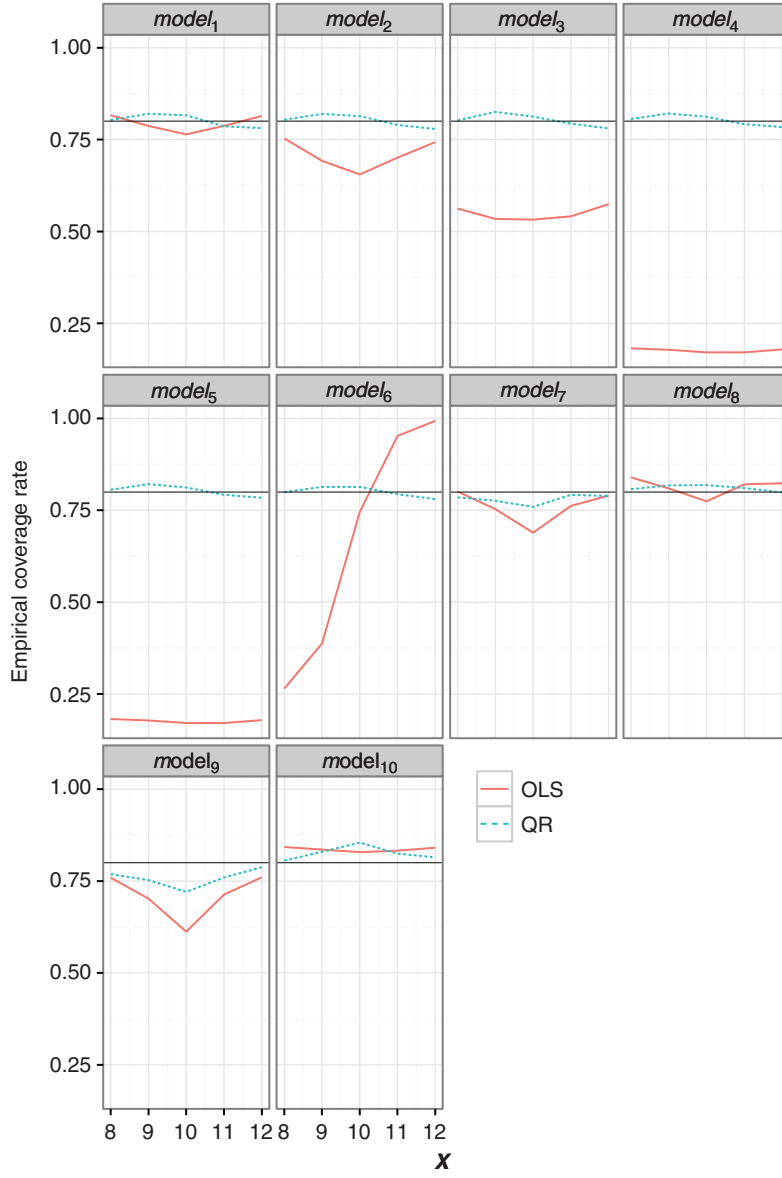

Figure 2.10 Empirical coverage levels computed on 1000 simulated datasets from the 10 different models (different panels) using the classical OLS prediction intervals (solid lines) at the level $(1-\alpha) = 80\%$*, and the QR prediction intervals (dashed lines) obtained using two regression quantile estimates* $[\hat{q}_Y(\theta = 0.1, X = x); \hat{q}_Y(\theta = 0.9, X = x)]$ *at five distinct values of* X *(abscissa). The horizontal line at* $y = 0.8$ *denotes the nominal coverage level.*

Table 2.6 Average interval width for OLS and QR prediction intervals computed using 1000 random samples extracted from each of the 10 considered models (rows of the table). The intervals are computed for five distinct values of X (columns of the table) to cover the whole range of the regressor.

		Average interval width [Nominal coverage level $(1-\alpha) = 80\%$]				
		$x = 8$	$x = 9$	$x = 10$	$x = 11$	$x = 12$
$model_1$	LS	2.63	2.59	2.57	2.59	2.63
	QR	2.51	2.52	2.53	2.55	2.56
$model_2$	LS	0.69	0.68	0.67	0.68	0.69
	QR	0.64	0.64	0.64	0.65	0.65
$model_3$	LS	1.57	1.55	1.54	1.55	1.57
	QR	1.35	1.36	1.36	1.37	1.37
$model_4$	LS	9.71	9.57	9.52	9.57	9.71
	QR	4.84	4.85	4.86	4.87	4.87
$model_5$	LS	9.71	9.57	9.52	9.57	9.71
	QR	4.84	4.85	4.86	4.87	4.87
$model_6$	LS	29.03	28.60	28.45	28.60	29.02
	QR	22.70	25.30	27.90	30.50	33.09
$model_7$	LS	2.69	2.65	2.64	2.65	2.69
	QR	2.61	2.61	2.60	2.60	2.59
$model_8$	LS	2.68	2.64	2.62	2.64	2.68
	QR	2.60	2.59	2.58	2.58	2.57
$model_9$	LS	3.06	3.01	3.00	3.01	3.06
	QR	2.98	2.97	2.96	2.95	2.94
$model_{10}$	LS	3.02	2.97	2.96	2.97	3.02
	QR	2.92	2.92	2.91	2.91	2.90

is referred to as the quantile process (Koenker 2005). By iterating the procedure illustrated in Section 2.1 to estimate $\beta(\theta)$ over the whole interval $(0, 1)$, it is then possible to obtain an estimate of the whole quantile process:

$$\{\hat{\beta}(\theta) : \theta \in (0, 1)\}.$$

In fact, QR estimates break the unit interval into a finite number of intervals, whose number and width is somewhat related to the dataset structure. The number of distinct regression quantiles is, indeed, related to the amount of data and to the number of variables in the sample, as we will show below using the same 10 models introduced in Section 2.2. Some considerations on the interval widths among subsequent distinct conditional quantiles are also discussed using a classical tool to compare inequality.

2.3.3.1 How to select the quantiles to estimate

In order to illustrate the quantile process, in the following we will focus on the slope estimates for the 10 illustrative models. A first comparison between the classical normal model ($model_1$) and the heteroskedastic model ($model_6$) is offered in Figure 2.11. On the x-axis, the different quantiles are depicted, while the estimates are shown on the y-axis: the two lines in Figures 2.11(a) and (b) refer to the two models. The presence of a heteroskedastic pattern in the error term for the $model_6$ permits an immediate appreciation for the QR strengths: the estimates for $model_1$ are practically constant over the different conditional quantiles, while the heteroskedastic pattern is effectively caught by the QR process. Although Figures 2.11(a) and (b) appear somehow different, they both depict the slope QR estimates for the same two samples. Figure 2.11(a) shows the QR estimates for the nine conditional deciles, that is estimating the conditional quantiles for $\theta = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The estimates for the nine conditional deciles are shown using points in the plot: the two lines joining the points have to be interpreted only as an attempt to approximate the real patterns for the conditional quantile function of the two models.

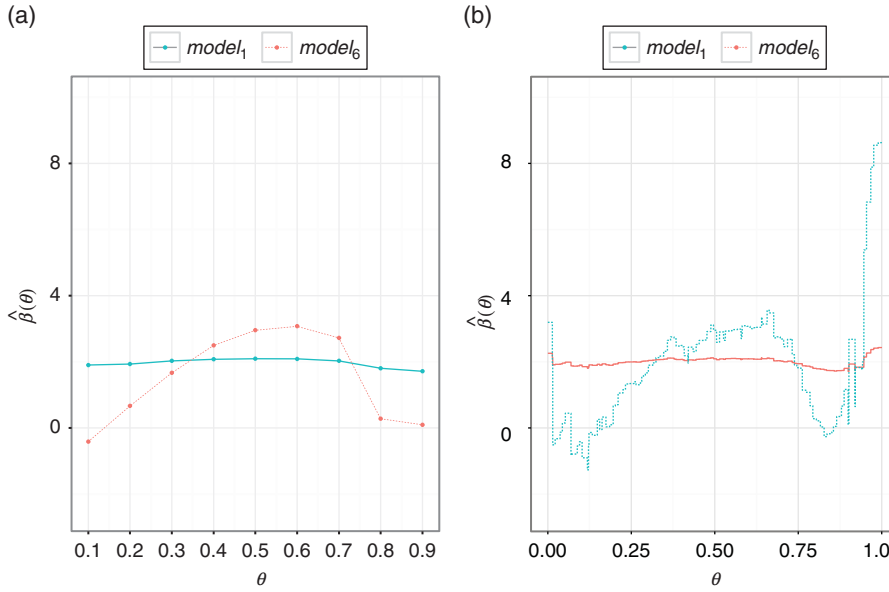

Figure 2.11 QR slope estimates for two random samples of n = *50 units extracted from* model$_1$ *(solid line) and* model$_6$ *(dashed line): (a) shows the estimates for the nine conditional deciles,* $\theta = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$*, while (b) shows the slope pattern in the case the whole quantile process is estimated. The conditional quantiles are represented on the* x*-axis and the coefficient values on the* y*-axis. In (b), subsequent estimates are connected by steps in order to stress that the quantile estimates are for intervals of quantiles.*

This is clearer by looking at the plot in Figure 2.11(b): here the estimates for the whole quantile process are depicted. In order to stress that QR estimates are for intervals of quantiles, subsequent estimates in Figure 2.11(b) are connected by steps. The two plots present a similarity in the underlying pattern, even if it is evident that only using an adequate number of quantiles is possible to appreciate all the details of the slope curve. The number of distinct QR solutions in this example consists of 59 quantiles and this explains the differences between the two plots.

A typical approach in real applications consists of using a dense grid of equally spaced quantiles in the unit interval $(0, 1)$, to obtain a fairly accurate approximation of the whole quantile process pattern. Again, referring to the same two samples, Figure 2.12(a) depicts the estimates computed in correspondence to 99 equally spaced conditional quantiles, that is the so-called conditional percentiles. The obtained pattern is in this case a fairly accurate reconstruction of the whole quantile process for both models, represented again in Figure 2.12(b). Obviously, in case the problem merits closer examination of a specific part or tail of the response conditional distribution, it is advisable to ask for a larger number of quantiles for that specific zone, in order to have a better view of the distribution.

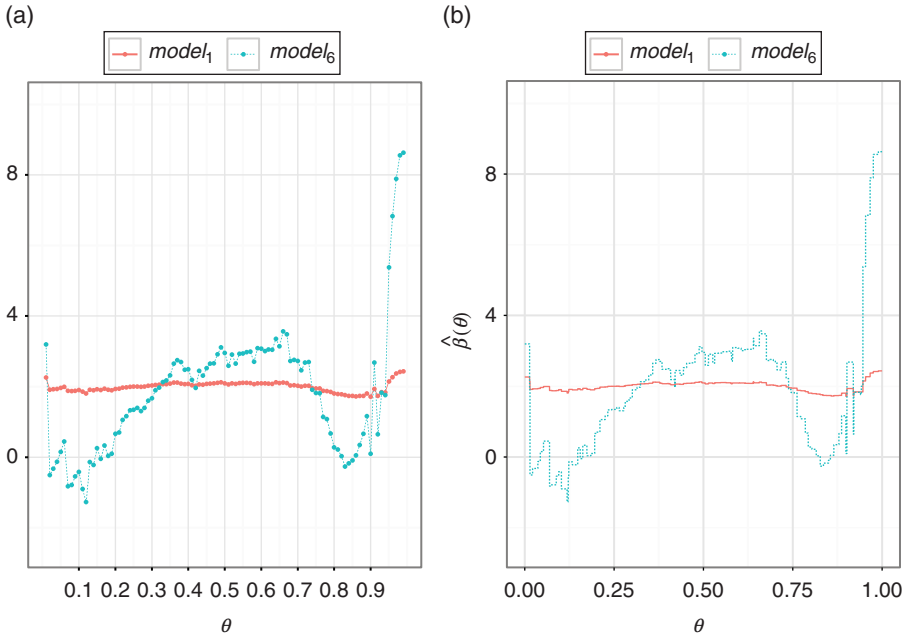

Figure 2.12 (a) QR slope estimates for 99 equally spaced conditional quantiles in the interval $(0, 1)$, that is the percentiles: the use of such a grid spanning the interval $(0, 1)$ allows obtaining a fairly accurate estimation of the whole quantile process (depicted again in (b)).

The whole quantile process for the four not normal homogeneous models (from $model_2$ up to $model_5$) is shown in Figure 2.13 and Figure 2.14, again for the slope coefficients. From Figure 2.13(a) follows that the presence of a higher degree of skewness ($model_4$) induces a higher variability in the slope estimates, in particular for the higher conditional quantiles. Figure 2.13(b) focuses only on the slope quantile process for $model_2$ and $model_3$ as they are not immediately evident in Figure 2.13(a) for the different scale of $model_4$. Moreover, it is interesting to note how the same error term, but with the opposite sign, implies a mirror pattern as depicted in Figure 2.14: the negative contribution of the error term to $model_5$ results in lower estimates for the left conditional quantiles, while adding the same error term ($model_4$) results in higher estimates for the upper part of the conditional distribution. Finally, the whole quantile process for the slopes of the four models with dependent errors (from $model_7$ up to $model_{10}$) is shown in Figure 2.15. For such models, the estimates for the two cases with negative autocorrelation coefficients ($model_7$ and $model_9$) are slightly higher with respect to the two models where the same values for ρ are used but with an opposite sign. Also in this case, in particular for the case $\rho = \pm 2$, the two quantile processes look specular.

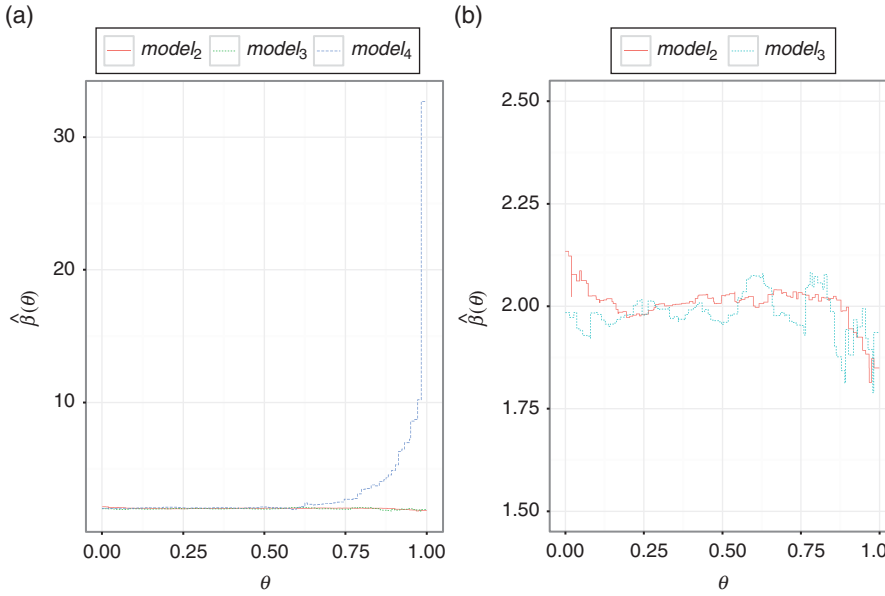

Figure 2.13 QR slope estimates for random samples of n = *50 units extracted from* model$_2$, model$_3$ *and* model$_4$ *(a) when the whole quantile process is estimated. The conditional quantiles are represented on the* x*-axis and the coefficient values on the* y*-axis. As the order of magnitude for* model$_4$ *is very high compared with the other two models, (b) is restricted to* model$_2$ *and* model$_3$ *estimates in order to better compare the corresponding patterns.*

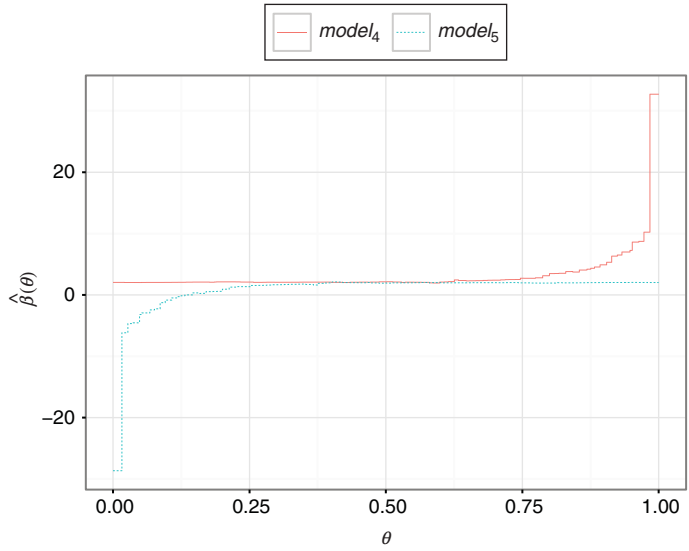

Figure 2.14 Slope quantile process for model$_4$ *and* model$_5$*: the different contribution of the same error term to the two models (positive in the case of* model$_4$ *and negative in the case of* model$_5$*) involves a mirror pattern for the QR estimates.*

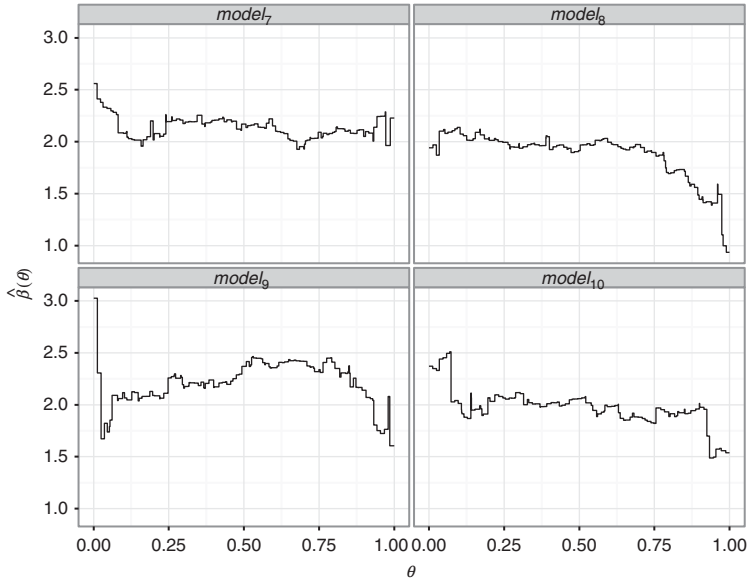

*Figure 2.15 Slope quantile process for the four dependent error models: the estimates for the two models with negative autocorrelation coefficients (*model$_7$ *and* model$_9$*) are slightly higher with respect to the two models with positive* ρ*. Moreover, the patterns for negative* ρ *are somewhat specular to the patterns for positive* ρ*, in particular for the cases* $\rho = \pm 0.2$*.*

2.3.3.2 How many distinct quantiles can be computed?

The distinct quantiles that can be computed is related both to the number of units and to the number of variables in the dataset.

The relationship with the number of units is obvious: it is sufficient to think to the difference in the stroke of the empirical distribution function by increasing the number of data, to guess that the number of possible solutions increases with the number of units. Figure 2.16 shows the number of distinct quantiles estimated for the 10 illustrative models when the sample size increases. The sample size is represented on the *x*-axis while the corresponding number of distinct quantiles is on the *y*-axis. The lines for the 10 models overlap showing no particular differences in terms of the type of model. The analysis has been carried out on 1000 samples starting from a sample size of 10 units up to a sample size of 10 000 units.

Figure 2.17 depicts the relation between the number of distinct solutions by taking into account jointly the number of units and the number of regressors in the model. The following data generating processes have been used:

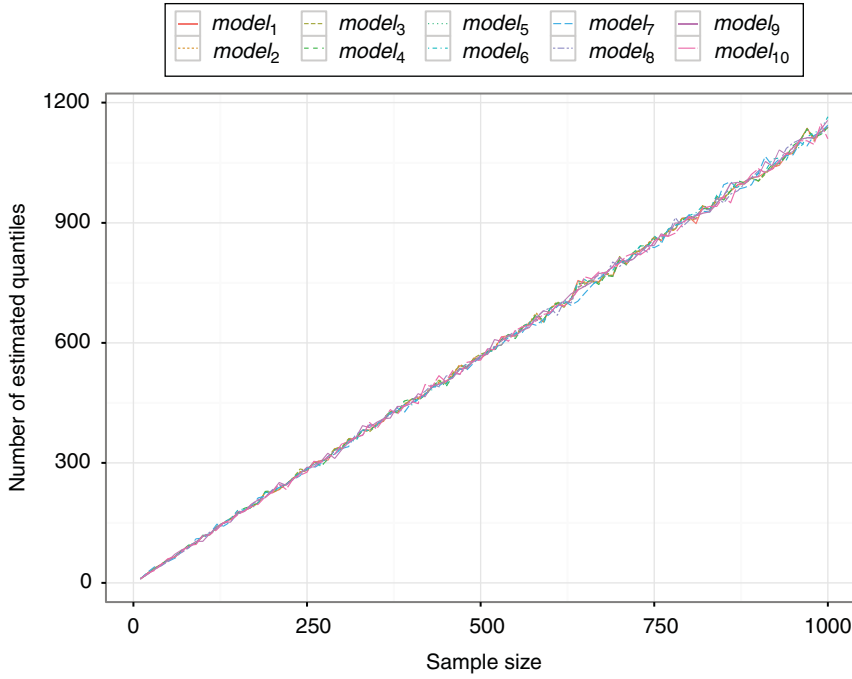

*Figure 2.16 Number of estimated distinct quantiles (*y*-axis) for the quantile process using different sample sizes (*x*-axis). The patterns for the 10 models introduced in Section 2.2 are shown using different lines: the number of estimated intervals increases when sample size increases but there are not substantial differences between the models.*

- a standardized normal error term $\mathbf{e} \sim N(0, 1)$;
- 10 standardized normal variables $X_i \sim N(0, 1), i = 1, \ldots, 10$;
- for each $i = 1, \ldots, 10$, the QR process has been estimated for the models:

$$\mathbf{y}^{(i)} = \sum_{j=1}^{i} 0.i \times X_i + \mathbf{e}$$

using $n = \{25, 50, 75, 100, 125, 150, 175, 200\}$ as sample sizes.

The results are represented through a dotplot where the number of distinct estimated quantiles is shown on the x-axis, the different sample sizes on the y-axis and

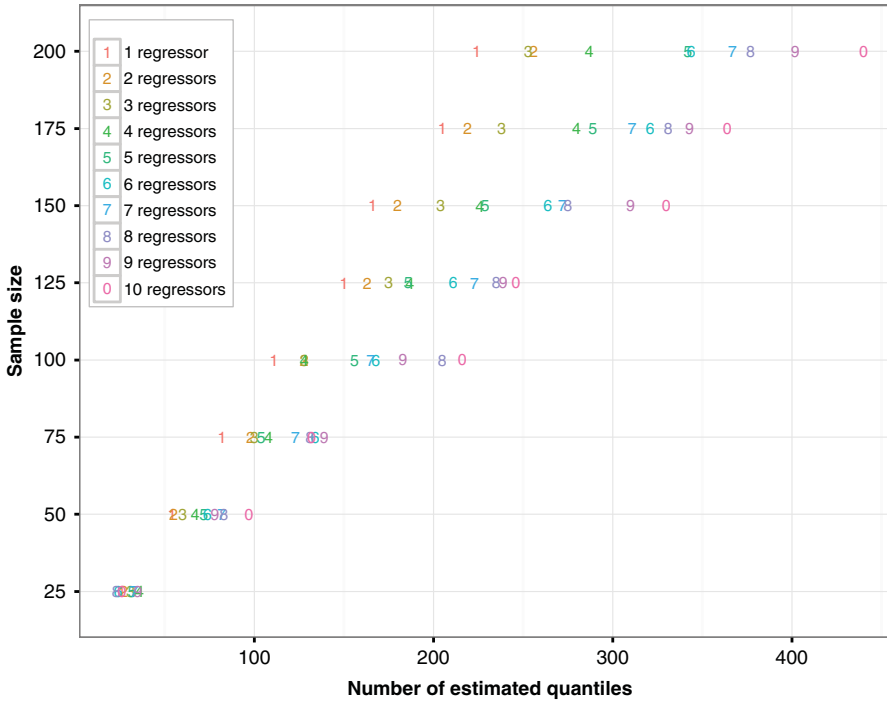

Figure 2.17 Number of estimated distinct quantiles (x-*axis) for the whole quantile process with respect to the sample size* (y-*axis) and the number of regressors used in the model. Data were generated using normal variates, both for the regressors and for the error term. The 10 models are represented using different points in the dotplot. When sample size increases, the number of estimated intervals increases (as already shown in Figure 2.16); the same happens increasing the number of variables in the dataset: differences in the number of estimated intervals are more pronounced for larger sample sizes.*

the various models, that is the effect of the different number of variables, as points in the plot. The use of a suitable labeling system allows to easily see the effect of the number of variables on the quantile process: in particular each model is represented using the corresponding number of regressors, where the label 0 depicts the model with 10 regressors. Starting from the bottom and moving to the upper part of the plot, the same pattern highlighted in Figure 2.16 is evident: following the trajectories of the same label on the different rows of the plot, it can be argued that the number of possible solutions increases with the number of units in the sample. Focusing instead on the lines of the plot, while for small sample sizes the number of estimated quantiles is not very different with respect to the number of used regressors, the range among the labels increases moving towards the upper part of the plot, denoting a more pronounced difference in the number of possible solutions.

Figure 2.18 illustrates the joint effect of the number of regressors and of the different error term. Using again normal variates as regressors, 10 further data generating processes have been used in order to take into account the different error terms introduced in Section 2.2. In particular, random sample from the 10 models summarized in Table 2.2 have been generated using 1, 5 and 10 regressors. Following the same line of Figure 2.17, the number of estimated distinct quantiles for the whole quantile process is depicted on the x-axis while the different sample sizes are on the y-axis. This time the different labels inside each panel refer to the 10 illustrative models: each model is represented using the corresponding number except for model_{10} whose label is 0. Finally the three panels correspond to the different number of involved regressors (1, 5, and 10).

The plot shows a substantial equivalence among the 10 models in the case of different numbers of regressors, as seen by comparing the three panels. The same pattern highlighted in Figure 2.16 and Figure 2.17 is evident also in this representation: when sample size increases the number of estimated intervals increases, but there are no substantial differences with respect to the number of regressors. Differences among the 10 models in the number of estimated intervals are more pronounced for larger sample sizes. The interesting pattern starts with sample sizes greater than 50 units, where the range among the labels starts to increase. The model characterized by the heteroskedastic error term ($model_6$, depicted by the label 6 on the plot) moves away from the others in the right direction. At the opposite side, for most of the sample size, the models with autocorrelated error (from $model_7$ up to $model_{10}$, depicted by the corresponding numbers on the plot). In some way this can be explained by the presence of a structure of dependence in the data that limits the number of different quantiles to estimate.

2.3.3.3 How distant are the possible distinct quantiles?

Finally, it is possible to show how the interval widths among the distinct possible solutions change both in terms of number of units and type of models. In particular, Figure 2.19 shows the different sample sizes on the y-axis and splits the interval (0, 1), depicted on the x-axis, in the case of random samples generated from $model_1$. The plot shows that estimates are quite scattered for small sample sizes, even if they

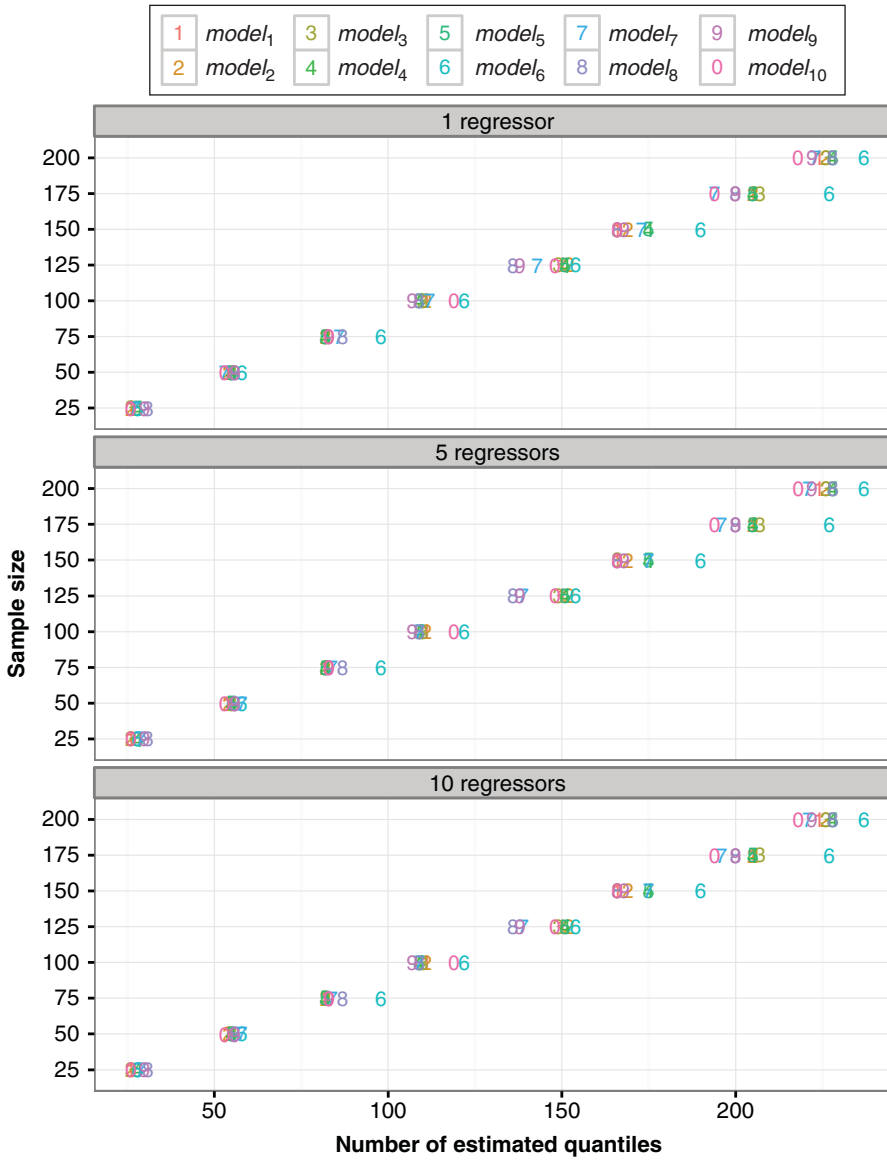

Figure 2.18 Number of estimated distinct quantiles (x-axis) for the whole quantile process. The sample size is depicted on the y-axis while the three panels refer to the different number of regressors in the model. Data were generated using normal variates while for the error terms the 10 models introduced in Section 2.2 (see Table 2.2) have been used. The 10 different obtained models are shown using different points in the dotplots.

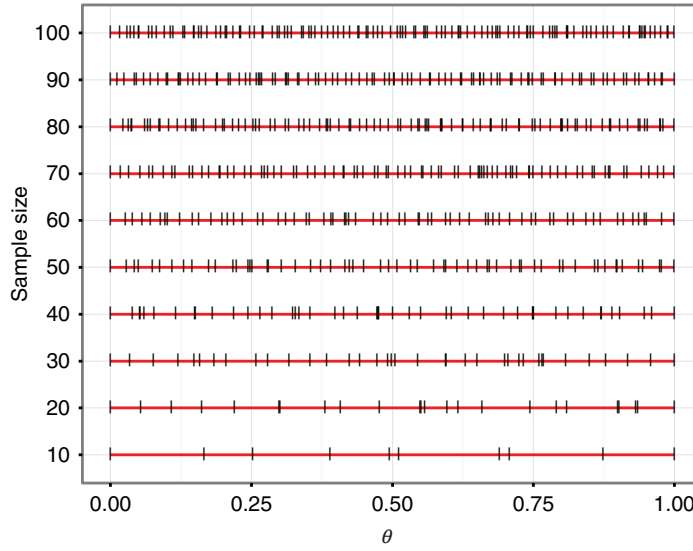

Figure 2.19 Estimated intervals for the whole quantile process for a random sample extracted from model$_1$ *with increasing sample sizes (*y*-axis). For each sample size the interval (0, 1), depicted on the* x*-axis, is split in unequal width intervals, identified by dashes on the corresponding unit length segments. The dashes depict the distinct quantiles that can be computed.*

tend to cover the whole unit interval starting from $n = 30$. Even if interval widths are different, increasing the number of possible solutions involves decreasing the average interval width, as one would expect.

The same representation is offered in Figure 2.20, comparing this time the quantile process among the 10 illustrative models (see again Table 2.2 for a recap). Figure 2.20(a) shows the intervals associated with the quantile process for a sample of size $n = 10$, while the case $n = 50$ is shown in Figure 2.20(b). Although some differences in the position of the conditional quantiles are evident in the case where $n = 10$, the 10 models are substantially equivalent both with respect to the number and position of the conditional quantiles using a sample of size $n = 50$. As already pointed out in Figure 2.16 and Figure 2.18, also with respect to the interval widths, the 10 models are substantially equivalent.

In order to further explore this aspect, it is useful to exploit an analysis typically used for inequality studies. Figure 2.21 shows the Lorenz curves (Lorenz 1905) for the interval widths for several random samples, one for each of the 10 illustrative models and for different sample sizes ($n = \{10, 50, 100, 200\}$): this plot confirms that the interval differences among distinct subsequent quantile solutions tend to become negligible as sample size increases for the models considered in the analysis. For the sake of interpretation it is worth remembering that the bisector plane corresponds to

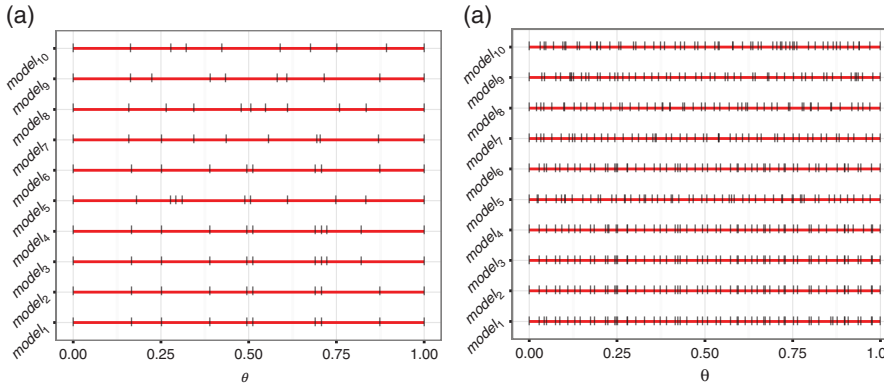

*Figure 2.20 Estimated intervals for the whole quantile process for 10 random samples extracted from the 10 models (see Table 2.2) considered (*y*-axis): (a) shows the intervals for random samples of size* n = *10; (b) shows the intervals for* n = *50.*

the case of perfect equality: the more the curves are close to the straight line, the more the corresponding intervals will tend to have the same width.

Although some of the plots shown in this chapter refer to single random sample(s), results do not substantially differ when the sampling procedure has been repeated: this is to say that conclusions are not strictly dependent on the particular sample(s).

2.4 Summary of key points

- The development and dissemination of QR started with the formulation of the QR problem as a linear programming 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 the case of datasets with a large number of observations and/or covariates, interior-point methods and/or a heuristic approaches permit estimates to be obtained.
- QR is an invaluable tool for facing heteroskedasticity, and provides a method for modeling the rates of change in the response variable at multiple points of the distribution when such rates of change are different. It is, however, also useful in the case of homogeneous regression models outside of the classical normal regression model, and in the case where the error independence assumption is violated, as no parametric distribution assumption is required for the error distribution.

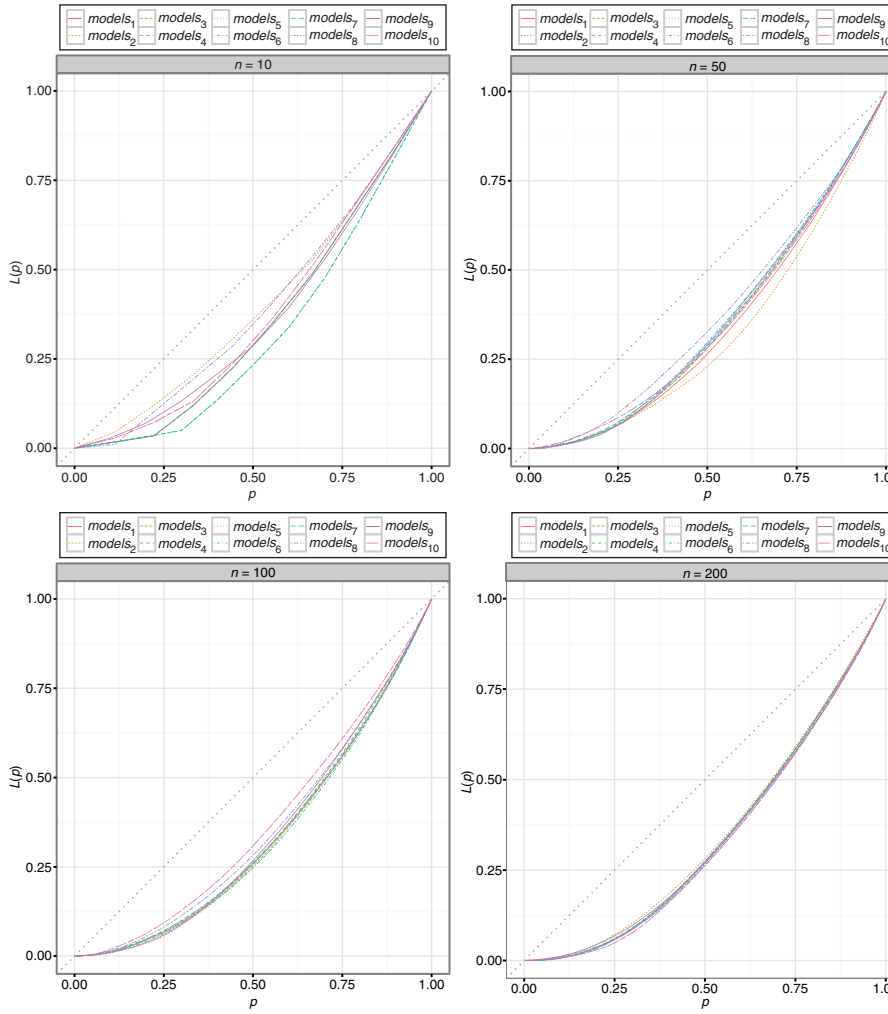

Figure 2.21 Lorenz curve for the width of the intervals of conditional quantiles when the whole quantile process is estimated: each panel refers to a different sample size, $n = \{10, 50, 100, 200\}$, while the different lines refer to the 10 illustrative models (see Table 2.2).

- Empirical conditional quantile function provides a standard distribution-free approach to obtain prediction intervals for population quantiles based on sample regression quantiles. Such use of regression quantile estimates has the advantage of robustness against departure from the normality assumption of the error terms.

- Although in many practical applications the focus is on estimating a subset of QR estimates, it is possible to obtain estimates over the whole interval (0, 1), the so-called quantile process.

- The number of distinct quantile solutions is related to the dataset structure, that is to the number of available units and variables. The interval (0, 1) is split into unequal width intervals: the number of distinct quantiles tends to cover the whole unit interval when the sample size increases, although the number of different quantiles is somewhat related to the nature of the error term.

References

Barrodale I and Roberts FDK 1974 An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis* **10**, 839–848.

Bloomfield P and Steiger W 1983 *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhauser.

Chen C 2007 A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics* **16**(1), 136–164.

Chen C 2004 An adaptive algorithm for quantile regression. In *Theory and Applications of Recent Robust Methods* (Hubert M, Pison G, Struyf A and Van Aelst S eds), 39–48. Birkhauser.

Chen C and Wei Y 2005 Computational issues for quantile regression. *Sankhya: The Indian Journal of Statistics* **67**(2), 399–417.

Dantzig G 1963 *Linear Programming and Extensions*. Princeton University Press.

Giloni A, Simonoff JS and Sengupta B 2006 Robust weighted LAD regression. *Computational Statistics & Data Analysis* **50**, 3124–3140.

Gujarati DN 2003 *Basic Econometrics*, International Edition. McGrawHill.

He X, Jureckova J, Koenker R and Portnoy S 1990 Tail behavior of regression estimators and their breakdown points. *Econometrics* **58**, 1195–1214.

Karmakar N 1984 A new polynomial time algorithm for linear programming. *Combinatorica* **4**, 373–395.

Koenker R 2005 *Quantile Regression*. Cambridge University Press.

Koenker R 2011 *quantreg: Quantile Regression*. R package version 4.76. http://CRAN.R-project.org/package=quantreg.

Koenker R and Basset G 1978 Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker RW and D'Orey V 1987 Algorithm AS 229: computing regression quantiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **36**(3), 383–393.

Lorenz MO 1905 Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* **9**(70), 209–219.

Matousek J and Gartner N 2007 *Understanding and Using Linear Programming*. Springer.

Neykov NM, Cizek P, Filzmoser P and Neytchev PN 2012 The least trimmed quantile regression. *Computational Statistics & Data Analysis* **56**, 1757–1770.

Portnoy S and Koenker R 1997 The Gaussian hare and the Laplacian tortoise: computation of squared-error vs. absolute-error estimators. *Statistical Science* **12**, 279–300.

Rousseeuw PJ and Hubert M 1999 Regression depth. *Journal of the American Statistical Association* **94**, 388–402.

SAS Institute Inc. 2010 *SAS/STAT 9.22 User's Guide. The QUANTREG Procedure*. SAS Institute Inc.

Zhou KQ and Portnoy SL 1996 Direct use of regression quantiles to construct confidence sets in linear models. *Annals of Statistics* **24**(1), 287–306.

Vanderbei RJ 2010 *Linear Programming. Foundations and Extensions*, 3rd Edition. Springer.

Wagner HM 1959 Linear programming techniques for regression analysis. *Journal of the American Statistical Association* **54**, 206–212.

Weisberg S 2005 *Applied Linear Regression*, 3rd Edition. John Wiley & Sons, Ltd.

3

Estimated coefficients and inference

Introduction

An introductory section shows the behavior of quantile regressions in datasets with different characteristics, adding more details to the discussion developed in Section 2.2. Section 3.2, using simulated data, shows the empirical distribution of the quantile regression estimator in the case of independent and identically distributed (i.i.d.) errors, non-identically distributed (i.ni.d.) errors and dependent (ni.i.d.) errors. The chapter then analyzes only the case of i.i.d. errors, while the other two cases are deferred to Chapter 5. Section 3.3 considers a small size real dataset and a very simple linear regression model where wages depend on education, to compare ordinary least squares and quantile regression estimates when the errors are i.i.d. Then the simple linear regression model is extended to comprise more than one explanatory variable, and elements such as *age*, *gender* and *type of work*, *dependent* or *independent*, *full time* or *part time*, are included. The tests considered in Section 3.4 allow to verify hypotheses on more than one coefficient at a time, in order to evaluate the validity of the selected explanatory variables. In the final specification, considering a very small dataset, wages turn out to depend upon age and degree of education.

3.1 Empirical distribution of the quantile regression estimator

In the linear regression model ordinary least squares (OLS) computes the line passing through the mean of the conditional distribution of the dependent variable, as in Figure 3.1, where the conditioning is with respect to the value assumed by the

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

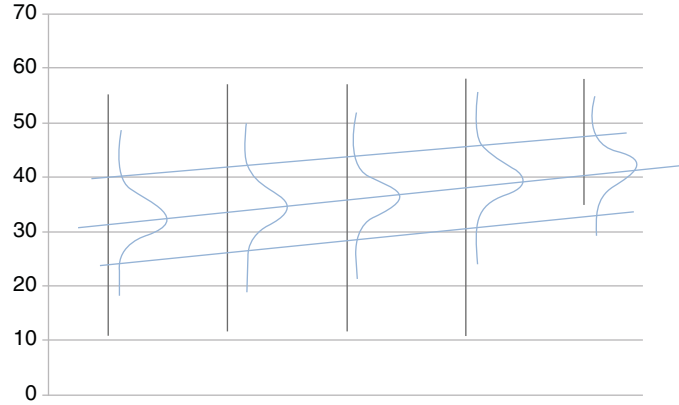

Figure 3.1 Independent normal error distributions, OLS regression, passing through the conditional means, the first and the third quartiles. These regression lines are equally spaced.

explanatory variable and the conditional distributions are depicted as independent normal distributions. Analogously, it is possible to compute regression lines passing through other measures of location of the conditional distribution of the dependent variable, for instance the mode or the quantiles. The latter are particularly useful in the case of asymmetric distributions since they provide measures of the proportionality among the regression variables at many different values/quantiles of the dependent variable, granting a deeper knowledge of the model under analysis. This is very informative if the dependent variable behaves differently from one tail to the other, such as in the case of skewed distributions. In Figure 3.2, for instance, the regression passing through the third quartiles of the conditional distributions, that is a line passing through the upper tails of the conditional distributions, is farther from the center than the line passing through the first quartiles. Thus the position of the fitted lines passing through different points of the conditional distributions allows a deeper knowledge of the model, and is particularly useful to point out asymmetry. OLS alone could not help since skewness can be assessed only through the comparison of different quantiles. In both Figure 3.1 and Figure 3.2 the quantile regressions (QRs) are parallel lines, which is the case when the conditional distributions are independent and identically distributed (i.i.d.). These figures plot the so-called location shift model, where changes among the different QRs show up only in the intercept. The case of nonparallel shifts in QRs, depicted in Figure 3.3, is characterized by non-identical conditional distributions (i.ni.d.): the error densities have the same shape but differ from one another in dispersion. This is the location and scale shift model, where QRs vary in both intercepts and slopes. Figure 3.3(a) shows the case of symmetric errors where as Figure 3.3(b) displays the case of skewed error distributions. The key characteristic is the different dispersion of the error densities. The QRs would have differing slopes even in the case of normal errors as long as their dispersion changes across the sample.

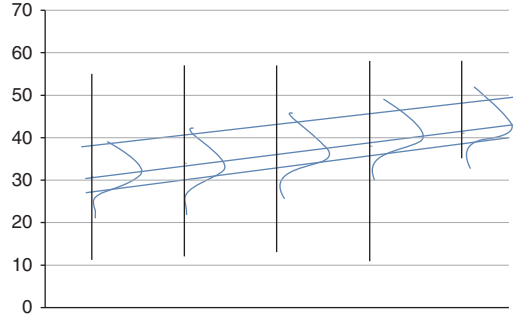

Figure 3.2 Asymmetrically distributed errors, regression lines passing through the first, second and third quartiles. The lines are parallel but unequally spaced due to the skewness of the conditional distributions.

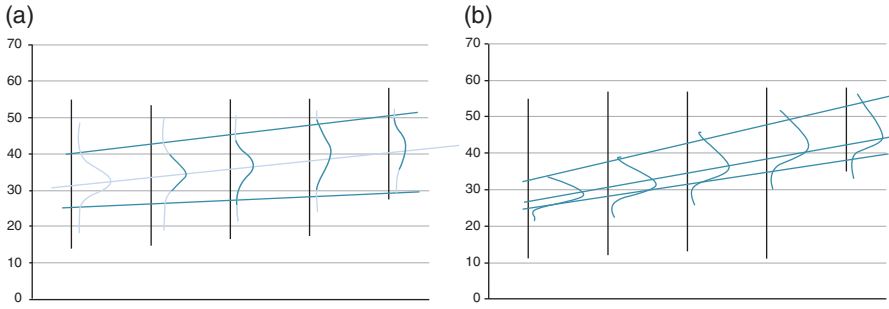

Figure 3.3 Symmetric (a) and asymmetric (b) non-identical conditional error distributions: median, first and third quartile regressions differing in slope.

Section 3.1.1 considers the i.i.d. case, while Section 3.1.2 looks at i.ni.d. data. The concluding part of Section 3.2 discusses how to pinpoint skewed conditional distributions.

3.1.1 The case of i.i.d. errors

In the linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ with i.i.d. errors having common strictly positive density f at the given quantile $f(F^{-1}(\theta)) > 0$, where $F^{-1}(\theta) = Q(\theta)$ is the quantile function defined in Equation (1.4), the quantile regression estimator $\widehat{\beta}(\theta)$ is asymptotically distributed as

$$\sqrt{n}[\widehat{\beta}(\theta) - \beta(\theta)] \to N(0, \omega^2(\theta)\mathbf{D}^{-1}), \tag{3.1}$$

with scale parameter $\omega^2(\theta) = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2}$ being a function of $s = 1/f(F^{-1}(\theta))$, the so-called sparsity function, $\mathbf{D} = lim_{n\to\infty} 1/n \sum_i \mathbf{x}_i^T \mathbf{x}_i$ being a positive definite matrix

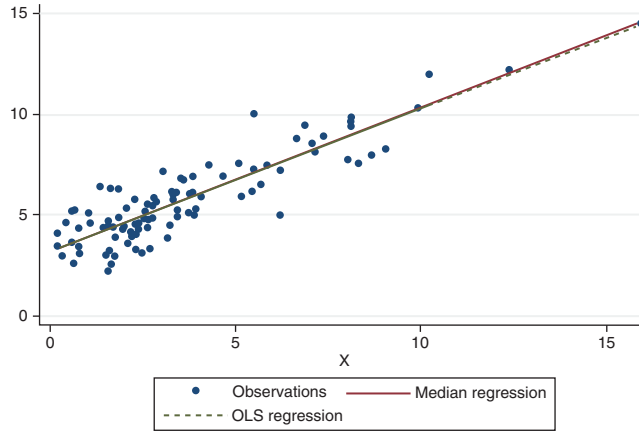

Figure 3.4 OLS and median regression when the errors are drawn from a standard normal distribution. The difference between the median and the OLS regression is minor.

and $\mathbf{x}_i$ the $(1, p)$ row vector comprising the i-th observation of the p explanatory variables (Koenker and Bassett 1978). In the above model $p = 2$ and $\mathbf{x}_i = [1\ x_i]$.

The asymptotic distribution of the QR estimator in (3.1) is explored by means of a small simulation experiment. One hundred observations of an independent variable, x_i, are drawn from a χ^2_4 distribution; the dependent variable is computed as $y_i = 3 + 0.7x_i + e_i$ and the error term e_i follows, in turn, a standard normal distribution, a χ^2_5 distribution independent of X and a Student-t with 2 degrees of freedom. The experiments with standard normal errors provide the OLS most favorable case, since under normality the OLS estimator coincides with the maximum likelihood estimator and thus yields the best linear unbiased (BLU) estimators. The χ^2_5 error distribution is introduced to see the behavior of the QR estimator with skewed distributions, while the Student-t is considered to model thick tails, that is distributions characterized by a larger probability of generating anomalous values. Each experiment is repeated 1000 times, in each replicate OLS and the QR model at the first, second and third quartile are estimated. Figure 3.4 presents the dependent and independent variables for the first of the 1000 iterations in the case of standard normal errors, along with the estimated OLS and median regressions. In this graph the mean (OLS) and median regressions overlap almost everywhere. Figure 3.5 presents one of the 1000 experiments with χ^2_5 errors. In this graph the dependent variable presents a number of quite large values, anomalous values in Y, which attract upward the OLS line, but not the median regression. Indeed the QR is robust to anomalous values in the dependent variable so that extreme values have a limited impact on the estimates, although this is not the case for outliers in the independent variable. Figure 3.6 considers one of the experiments with a Student-t error distribution. This distribution, which assigns a large probability to the tails, yields even more outlying observations and once again the OLS line is attracted by the outlying values.

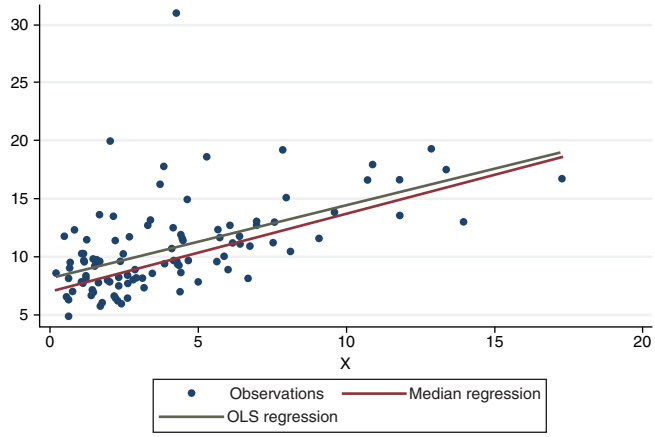

Figure 3.5 OLS and median regression when the errors are drawn from a χ^2_5 distribution. This error distribution generates five outliers in the dependent variable, as can be seen in the top section of the graph. The outliers shift upward the OLS fitted line but not the median regression.

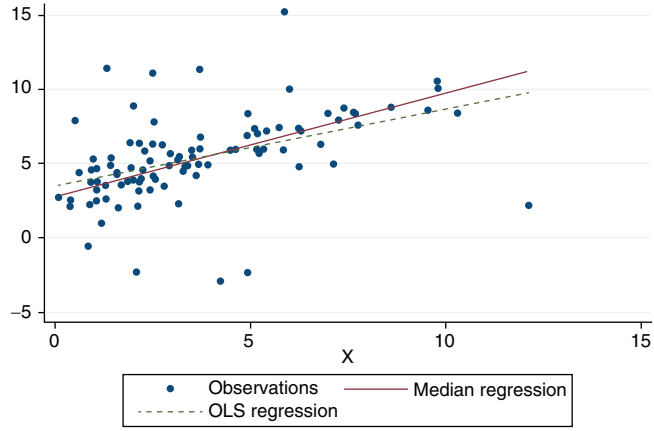

Figure 3.6 OLS and median regression when the errors are drawn from a Student-t_2 distribution. This distribution generates outliers which tilt the OLS fitted line but not the median regression.

Figure 3.4, Figure 3.5 and Figure 3.6 show the results computed in only one iteration. By repeating 1000 times each experiment, the estimated coefficients of each iteration are collected and the empirical distributions of the estimated slope are reported in Figure 3.7, Figure 3.8 and Figure 3.9, while the summary statistics of these distributions are in Table 3.1. Figure 3.7 presents the empirical distributions of the slope coefficient as estimated at the first quartile, the median and the last quartile

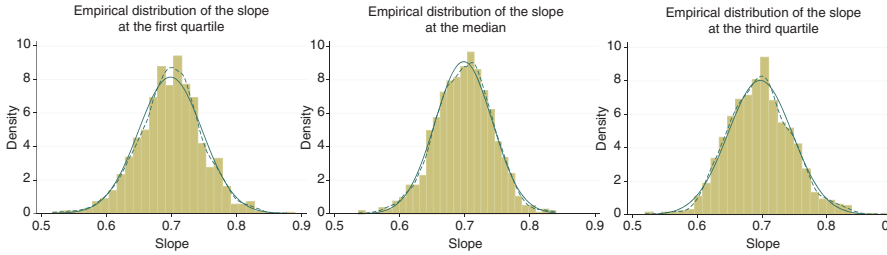

Figure 3.7 Empirical distributions of the QR estimates of the slope in 1000 replicates in the case of i.i.d. standard normal errors. The solid line is the normal density and the dashed line is the Epanechnikov kernel density. The good approximation of the empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.

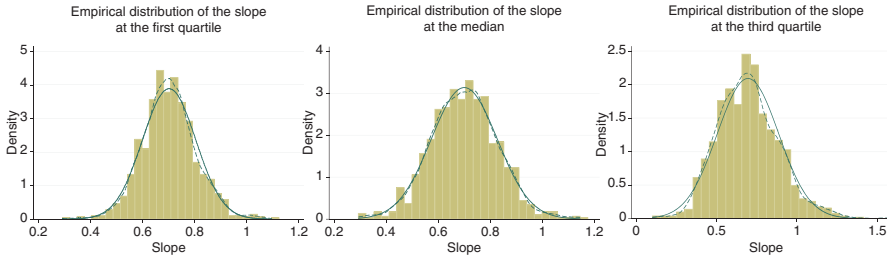

Figure 3.8 Empirical distributions of the QR estimates of the slope in 1000 replicates in the case of i.i.d. errors following a χ^2_5 distribution. The solid line is the normal density and the dashed line is the Epanechnikov kernel density. The good approximation of the empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.

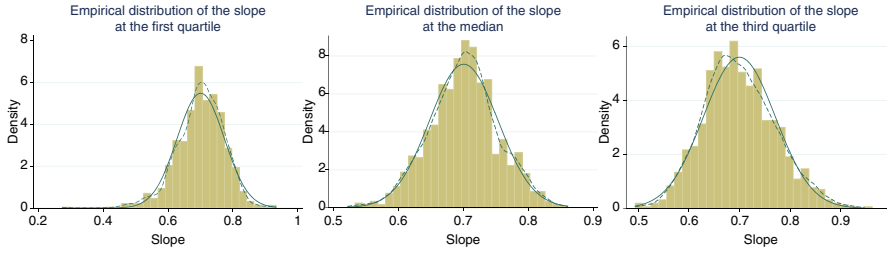

Figure 3.9 Empirical distributions of the quantile regression estimated slope in 1000 replicates in the case of i.i.d. errors following a Student-t with 2 degrees of freedom. The solid line is the normal density and the dashed line is the Epanechnikov kernel density. The good approximation of the empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.

Table 3.1 Empirical distribution of the estimated slope in 1000 replicates, with $N(0,1)$, χ^2_5, t_2, non-identically distributed, and dependent errors, in the model $y_i = 3 + 0.7x_i + e_i$. When the errors are normally distributed OLS yields greater precision, while with non-normal errors the dispersions of the OLS empirical distributions is greater than the dispersion of the empirical distributions of the QR estimates.

Error		Estimator			
		$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	OLS
$e_i \sim N(0,1)$	Mean	0.698	0.698	0.697	0.698
	Standard deviation	0.049	0.043	0.049	**0.036**
$e_i \sim \chi^2_5$	Mean	0.702	0.699	0.695	0.699
	Standard deviation	**0.102**	0.127	0.190	0.117
$e_i \sim t_2$	Mean	0.700	0.700	0.699	0.698
	Standard deviation	0.073	**0.052**	0.071	0.111
$e_i \sim \begin{cases} N(0,1), & \text{if } i = 1,50 \\ N(0,100), & \text{if } i = 51,100 \end{cases}$	Mean	0.661	0.703	0.754	0.707
	Standard deviation	0.272	**0.092**	0.282	0.256
$e_i = 0.5e_{i-1} + a_i$	Mean	0.703	0.695	0.701	0.699
	Standard deviation	**0.051**	0.052	0.055	0.700

The smaller standard deviations within each group of experiments are given in bold.

regression, along with the corresponding kernel (Epanechnikov 1969) and normal density. Then the empirical distributions of the slope coefficient when the error term follows a χ^2_5 distribution independently drawn from X are presented in Figure 3.8. Finally Figure 3.9 presents the empirical distributions when the errors are drawn from a Student-t_2 distribution. The approximation of the empirical distributions of the slope, as estimated by the QR, to the normal is quite evident in each figure, no matter what quantile is estimated and what error distribution is implemented, normal, skewed or thick tailed. These empirical distributions are unimodal, symmetric, and always centred on the true value $\beta_1 = 0.7$. The first two rows of Table 3.1 present sample mean and standard deviation of the empirical distributions of the slope coefficient estimated at the first, second, third quartile and at the mean regression, by OLS, when the error distribution is a standard normal. The second group of two rows summarizes the experiments with skewed error distributions and the third group considers the case of Student-t errors. These results confirm that the QR estimator is indeed unbiased, since the sample mean is equal to the true coefficient or is very close to it. In the case of normality OLS provides the smallest standard deviations, as expected. With non-normal errors the standard deviations are generally larger, indeed in these experiments the empirical distributions are more dispersed than in the standard normal case. Thus skewed and thick tails errors imply lower precision in the

estimates. While in the normal error experiment OLS is more precise than the QR estimator, with non-normality the precision of QR improves upon OLS. The OLS empirical distributions have a quite large standard deviation particularly in the experiments with Student-t errors. In the QR, however, the standard deviation generally increases in moving away from the median, particularly at the extreme quantiles. This is due to a reduced density of the data in the tails.

3.1.2 The case of i.ni.d. errors

Non-identically distributed errors are generally characterized by changing variance across the sample, which implies an error density f_i changing in the sample. A typical example is given by consumption expressed as a function of income. When income is high there are many possible alternative combinations of saving and spending, while with low incomes the choices between spending and saving are more limited. This causes the variability of consumption to increase with income. Figure 3.10 presents a similar case. In the graph changes in consumption, $\Delta c_t = c_t - c_{t-1}$, are plotted over time for a sample of size $n = 124$. It is quite evident that at the beginning of the sample the dispersion of the data is much lower than at the end of the sample. These are quarterly data on consumption in Italy, where c_t is observed from the first quarter of 1980 to the first quarter of 2011.[1] By splitting the sample exactly in two halves, in the first subset of 62 observations for the period 1980–1995, the standard deviation of Δc_t is $std(\Delta c_t) = 941$, while in the second half of the sample, from

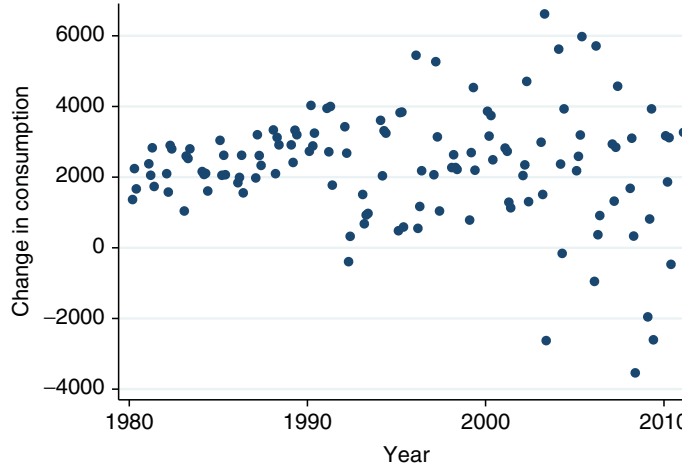

Figure 3.10 Italian data on changes in consumption, sample size n = *124, quarterly data from 1980 to 2011. The dispersion at the beginning of the sample is sizeably smaller than at the end of the sample. Data in the 1980s are realizations of* $\boldsymbol{f}_{\mathrm{i}}$ *having smaller variances than the* $\boldsymbol{f}_{\mathrm{i}}$ *generating the observations in the last decade.*

[1] Source: OECD Quarterly National Account.

1995 to 2001, it becomes $std(\Delta c_t) = 2061$. Once again there is ample evidence of non-identically distributed densities. While the analysis of i.ni.d. models is deferred to Chapter 5, here the empirical distribution of the QR estimator is considered. In the case of non-identically distributed f_i, the asymptotic distribution of the QR estimator is given by

$$\sqrt{n}[\widehat{\beta}(\theta) - \beta(\theta)] \to N(0, \theta(1-\theta)\mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1}), \tag{3.2}$$

where $\mathbf{D}_1(\theta) = lim_{n\to\infty} 1/n \sum_i f_i(F^{-1}(\theta))\mathbf{x}_i^T\mathbf{x}_i$ is a positive definite matrix (Koenker and Basset 1982a). Equation (3.2) collapses into Equation (3.1) in case of i.i.d. errors. Indeed the covariance matrix

$$\theta(1-\theta)\mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1}$$

simplifies in $\omega^2\mathbf{D}^{-1}$, with

$$\omega^2(\theta) = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2}.$$

This occurs since $f_i(F^{-1}(\theta)) = f(F^{-1}(\theta))$ is constant and can be singled out of $\mathbf{D}_1(\theta)$. As a consequence $\mathbf{D}_1(\theta)$ and $\mathbf{D}$ coincide and the term $\mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1}$ collapses to $\mathbf{D}^{-1}$.

The simulations for i.ni.d. errors consider the same explanatory variable $X \sim \chi_4^2$ but a different definition of the error term e_i. In each iteration, the first 50 observations of e_i are drawn from a standard normal while the second half of the sample follows a zero mean normal distribution having variance 100. In this subset the errors, and thus the y's, are more dispersed than in the first half of the sample. Once again the dependent variable is defined as $y_i = 3 + 0.7x_i + e_i$.

Figure 3.11 presents the dependent and independent variables for the first of the 1000 iterations in the case of i.ni.d. errors, along with the estimated OLS and median regressions. Their slopes are somewhat different, but one single experiment does not lead to any conclusion. In 1000 replicates the QR estimated values of the slope coefficient are collected and their empirical distributions are reported in Figure 3.12. The graphs show that the empirical distributions of the estimated slopes are centred around the true value, so that the slope estimator is once again unbiased, but away from the median the dispersion is greater and the empirical distributions become skewed. The second to last group of two rows in Table 3.1 reports sample mean and standard deviation of the empirical distributions in the case of i.ni.d. errors. The dispersion of the OLS empirical distribution more than doubles the dispersion of the estimated slope at the median regression.

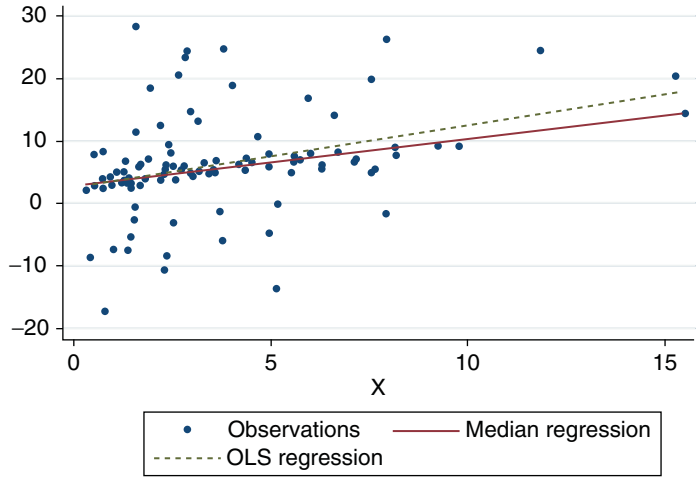

Figure 3.11 OLS and median regression in the case of i.ni.d. errors. This error distribution generates half the observations by a standard normal, and these are the points closer to the fitted line. The remaining half of the data are generated by a $\mathrm{N}(0, 100)$*, and these are the farthest observations in the graph. The mean and median regression have differing slopes.*

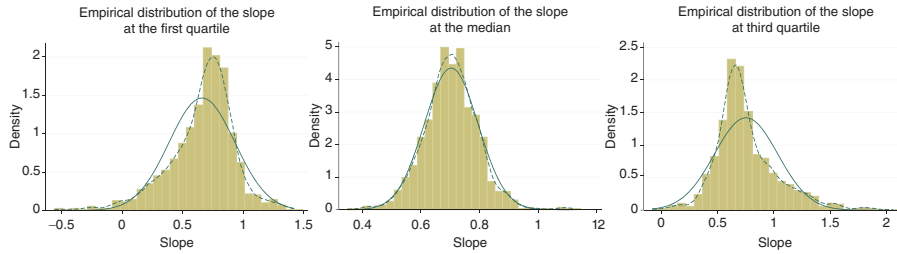

Figure 3.12 Empirical distributions of the QR estimated slope in 1000 replicates in the case of i.ni.d. errors. The solid line is the normal density and the dashed line is the Epanechnikov kernel density. Away from the median the distributions become skewed, left skewed at the lower quartile and right skewed at the higher one. At the median the approximation of the empirical distribution and its kernel smoothed version to the normal density is good.

3.1.3 The case of dependent errors

Dependent error distributions occur mostly with time series data, where dependence implies that recent errors are influenced by their own previous values. The density f_i is the same across the sample but is no longer independent from its own past values

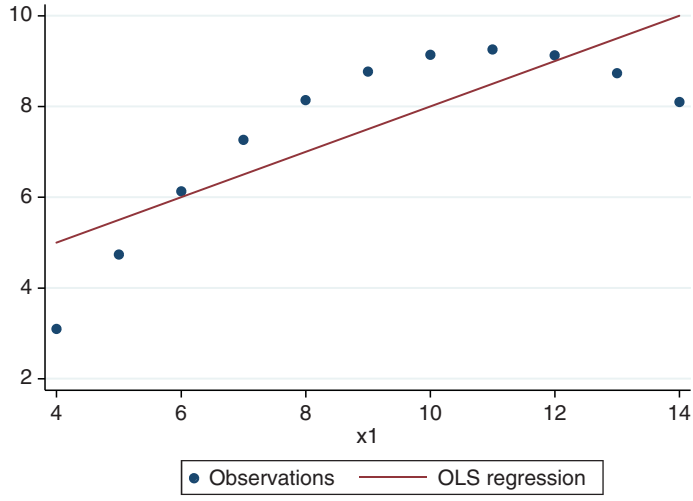

Figure 3.13 Erroneous linear specification of the regression model.

f_{i-h}. The simplest case defines the errors of the selected linear regression model, $y_i = \beta_0 + \beta_1 x_i + e_i$, with $e_i = ae_{i-1} + \mathrm{a}_i$ where a_i is an i.i.d. error term. In this model the i-th observation of the dependent variable is influenced by the previous value of the error, e_{i-1}. This is a first order serial correlation model, $AR(1)$, but it can easily be generalized to include more lags in the definition of e_i, as for instance in the case of an $AR(3)$ process where $e_i = a_1 e_{i-1} + a_2 e_{i-2} + a_3 e_{i-3} + \mathrm{a}_i$. The general case is an $AR(q)$ process where q past errors, from e_{i-1} up to e_{i-q}, influence y_i, $e_i = \sum_{h=1}^{q} a_h e_{i-h} + \mathrm{a}_i$.

However, besides models where past errors influence the actual value of the dependent variable, serial correlation can also be the result of poor model specification. Figure 3.13 presents such a case, with a nonlinear model incorrectly estimated by a linear equation. Figure 3.14 reports the corresponding residuals, fluctuating around zero and linked to their own previous values. Indeed the estimate of the correlation between these residuals and their one period ahead values yields a non-zero result, $corr(\widehat{e}_i, \widehat{e}_{i-1}) = -0.168$.

This section focuses on $AR(1)$ errors leaving further considerations on the ni.i.d. models to Chapter 5. The asymptotic distribution of the median regression estimator in the case of $AR(1)$ errors is

$$\sqrt{n}[\widehat{\beta}(\theta) - \beta(\theta)] \rightarrow N(0, \omega^2(\theta)\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}), \tag{3.3}$$

where $\mathbf{A} = lim_{n\rightarrow\infty}\mathbf{D} + 1/n \sum_i \psi(e_i)\psi(e_{i-1})(\mathbf{x}_i^T\mathbf{x}_{i-1} + \mathbf{x}_{i-1}^T\mathbf{x}_i)$ is a positive definite matrix, $\sum_i \psi(e_i)\mathbf{x}_i^T = \sum_i sgn(e_i)\mathbf{x}_i^T$ is the gradient of the QR and $\psi(\ )$ is the derivative of the QR objective function and is equal to the *sign* function (Weiss, 1990).

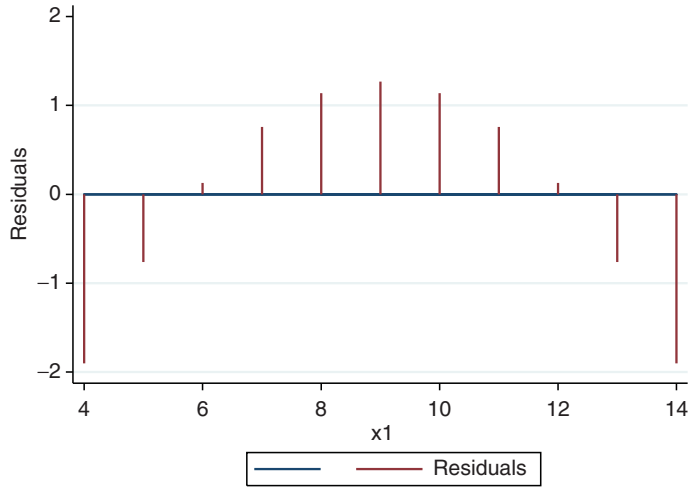

Figure 3.14 Serially correlated residuals due to an erroneous linear specification of the estimated model.

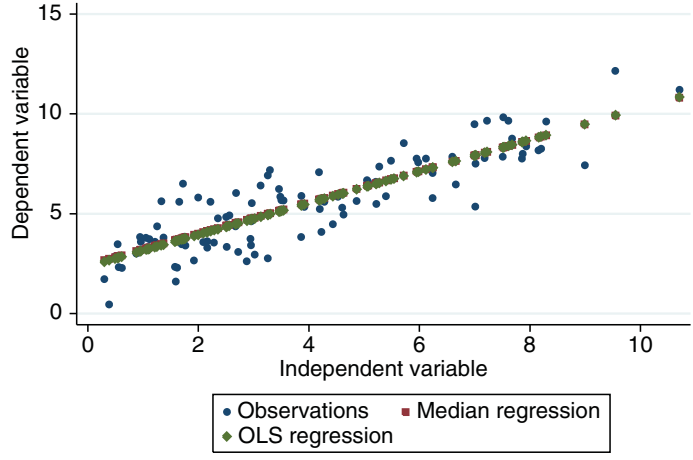

Figure 3.15 OLS and median regression in the case of dependent errors. In this example recent observations depend upon previous errors. In this model the errors are AR(1) and e_{i-1} influences y_i.

Figure 3.15 presents the dependent and independent variables for the first of the 1000 iterations in the case of $AR(1)$ errors, along with the estimates of OLS and median regressions which coincide.

In the simulations of this set of experiments the linear regression model is, as usual, $y_i = 3+0.7x_i+e_i$, while the error term is generated as $e_i = 0.5e_{i-1}+a_i$, a_i being an i.i.d. standard normal. The sample mean and standard deviation of the empirical

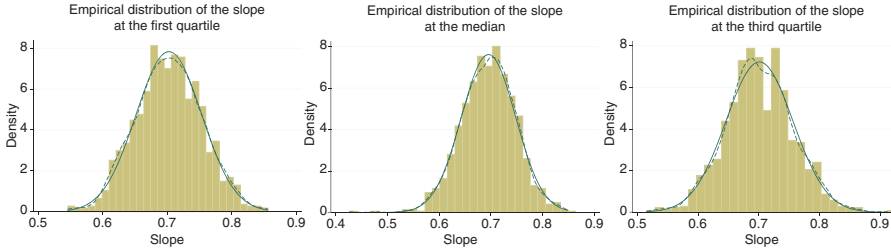

Figure 3.16 Empirical distributions of the QR estimated slope in 1000 replicates in the case of dependent errors. The solid line is the normal density and the dashed line is the Epanechnikov kernel density. The good approximation of the empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.

distributions are reported in the last tow rows of Table 3.1. The dispersion of the OLS estimates is sizeably larger than the dispersion of all the empirical distributions of the QRs. All the distributions are centred on the true slope coefficient. Figure 3.16 shows symmetric and bell shaped distributions at all quantiles.

3.2 Inference in QR, the i.i.d. case

This section considers two examples, a real dataset and a simulated one. The real dataset is a small sample comprising log of *wages* and *degree of education* in 2006 in Italy.[2] It is a very narrow subset, $n = 49$, extracted from one of the 11 waves of the Survey of Household Income and Wealth provided by Banca d'Italia. The selected equation is $lwage_i = \beta_0 + \beta_1 education_i + e_i$ and measures the existing link between the log of wages, $lwage_i$, and education. The dependent variable is the logarithm of hourly wage net of taxes and social security contribution, expressed in euros using the year 2000 consumer price index deflator. The explanatory variable is education which takes the values 8, 13, and 18 according to the completed education level respectively, junior high, high school, and university. The model assumes that wage grows – or decreases in the unfortunate case of a negative slope coefficient – as a function of the attained education level. The model is overthrown in the case the null hypothesis for the slope, $H_0 : \beta_1 = 0$, is not rejected.

The second example considers an artificial dataset characterized by a larger sample of size $n = 100$. It is one of the 1000 simulated datasets considered in the previous section to describe the empirical distributions of the QR estimated coefficient in case

[2] The English version of the questionnaire and the full dataset can be downloaded from the Banca d'Italia website at http://www.bancaditalia.it/statistiche/indcamp/bilfait/dismicro/annuale. The survey took place in 1989, 1991, 1993, 1995, 1998, 2000, 2002, 2004, 2006, 2008 and 2010. The full sample for 2006 is very large, but most observations have been excluded in order to have a small and manageable i.i.d. sample.

of i.i.d. skewed errors. The data generating process is $y_i = 3 + 0.7x_i + e_i$, x_i is drawn from a χ^2_4 distribution and the error term e_i follows a χ^2_5 independently drawn from X. This is to investigate the behavior of QR in a controlled setting where it is well known that the conditional distribution is asymmetric. In contrast, in the case of real data, little is known about the characteristics of the conditional distribution. Table 3.2 presents the summary statistics for the dependent variable of both datasets. In both cases the mean is larger than the median. This signals a positive asymmetry, indeed the skewness coefficients are positive. The Y variable is also characterized by a greater dispersion. Figure 3.17(a) presents the density of the *lwage* variable, along with its kernel smoothing, and the normal density. Figure 3.17(b) shows the empirical quantile function, $Q(\theta) = F^{-1}(\theta)$ in comparison with the normal. Analogously Figure 3.18 presents the density and the quantile function of the simulated dependent variable Y. Both figures present an asymmetric density function; the dispersion in y_i is larger, and its density in the tails is greater than normal, as can be seen in Figure 3.18(b), which shows a sizeable divergence from the normal quantile function in the tails.

Then the analysis moves to the regression model. Figure 3.19 presents the scatter plot of the wage data, which are gathered for the three levels of education considered: junior high, high school, and university. The graph reports the median and the OLS estimated regressions, which measure the impact of education, respectively, at the conditional median and at the conditional mean of the dependent variable, log

Table 3.2 Summary statistics of the dependent variables of the two examples.

	Mean	Median	Standard deviation	25th quantile	75th quantile	Skewness
lwage	2.322	2.225	0.530	1.937	2.785	0.475
Y	9.968	9.657	2.927	7.821	11.732	0.365

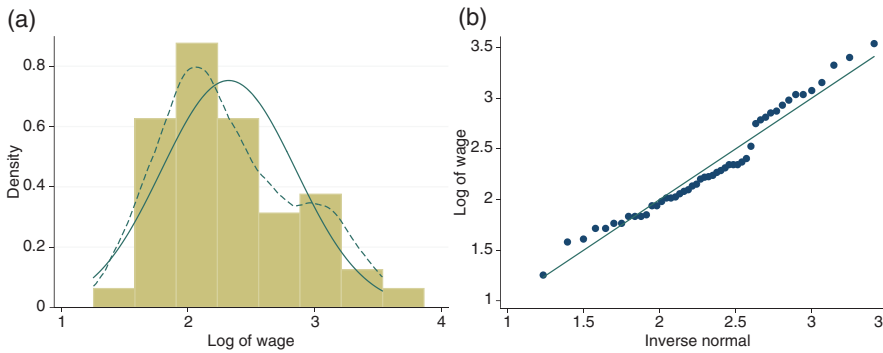

Figure 3.17 (a) The density of log of wage. The solid line is the normal density while the dashed line is the Epanechnikov kernel density. (b) The empirical quantile function of log of wage compared with the normal analog, represented by the straight line.

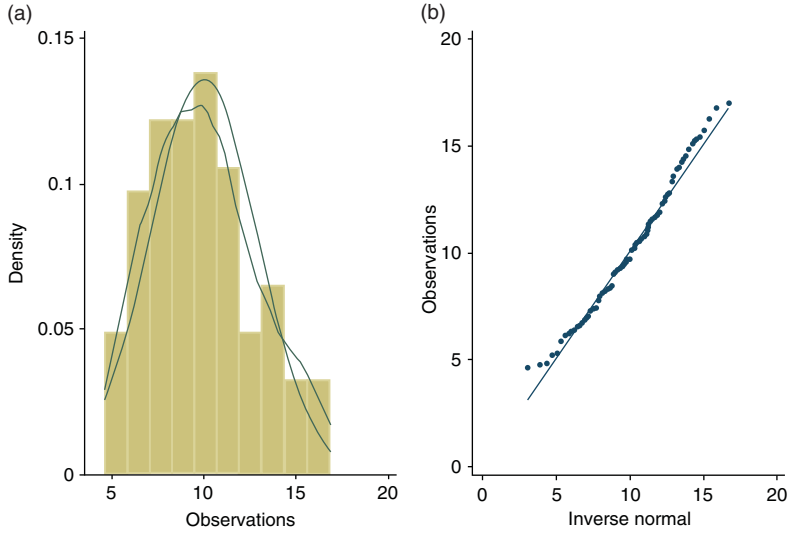

Figure 3.18 (a) The density of Y. The solid line is the normal density while the dashed line is the Epanechnikov kernel density. (b) The empirical quantile function of Y compared with the normal analog, represented by the straight line.

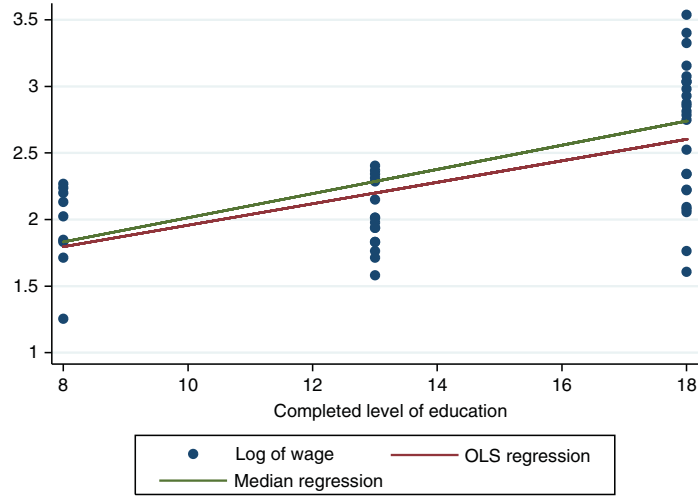

Figure 3.19 OLS and median estimated regression of the wage equation. Sample size n = 49. *The two lines have similar intercept and differing slopes, however the QR estimates do not statistically differ from the OLS results.*

Table 3.3 OLS and QR estimated coefficients at five different quantiles, wage equation.

	Estimator					
	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	$\theta = 0.90$	OLS
education	0.068	0.044	0.091	0.090	0.091	0.081
	(0.033)	(0.016)	(0.029)	(0.026)	(0.032)	(0.016)
constant	0.828	1.357	1.107	1.299	1.505	1.149
	(0.508)*	(0.247)	(0.430)	(0.391)	(0.461)	(0.245)

Standard errors are given in parentheses, sample size $n = 49$. The asterisk indicates a coefficient having low Student-t statistic.

of wage. The OLS estimated regression is an upward sloped line, with a slope of $(\widehat{\beta}_1) = 0.081$ and a Student-t statistic equal to $t(\widehat{\beta}_1)$=5.74. The slope is statistically different from zero and provides evidence in favor of wage being directly proportional to education. Additional education provides, on average, an 8% increase in log of wage. These results are reported in the last column of Table 3.3.

The conditional mean model can be compared with the estimates at the median regression. The median regression is the robust counterpart of the OLS regression exactly as the median is the robust measure of location complementing the mean. In Figure 3.19 the median regression for the wage equation is steeper than the OLS line, but it turns out that the difference is not statistically relevant. Indeed the estimated slope at the median regression, $\widehat{\beta}_1(.50) = 0.091$, falls within the $\pm 2\widehat{se}(\widehat{\beta}_1)$ OLS confidence interval, given by $c.i.(\widehat{\beta}_1)_{OLS} = [0.0481; 0.113]$. The statistical relevance of the median estimated regression is assessed through the usual Student-t statistic (Koenker 2005). Indeed, since the QR estimator is asymptotically normal, standardizing it with its estimated standard error in place of the unknown true standard error yields a Student-t distribution. Under the null $H_0 : \beta(\theta) = 0$, the Student-t test is computed by the ratio between the estimated coefficient and its own estimated standard error, just as occurs in the OLS regression model. For the slope the t-test is $t(\widehat{\beta}_1(.50)) = \widehat{\beta}_1(.50)/\widehat{se}(\widehat{\beta}_1(.50)) = 3.13$ and the null is rejected. For the intercept the t-test is $t(\widehat{\beta}_0(.50)) = \widehat{\beta}_0(.50)/\widehat{se}(\widehat{\beta}_0(.50)) = 2.57$ and the null is again rejected. Figure 3.19 shows that the estimated slopes at the conditional median and at the conditional mean are different, while the two estimated intercepts are very close to one another in this example.

Figure 3.20 presents the OLS and the median regression for the simulated data. In this graph the OLS estimated line moves upward due to a greater dispersion of the observations. The thick-tailed density generates a number of observations quite far from the center of Y. These attract the OLS fitted line causing a change in the intercept but not in the slope. The figure shows that this example is characterized by differing intercepts and similar slopes. The estimated coefficients are all significantly different from zero. In particular the OLS slope is $\widehat{\beta}_1 = 0.587$, not much different from the slope of the median regression, $\widehat{\beta}_1(0.50) = 0.629$, which falls within the $\pm 2\widehat{se}(\widehat{\beta}_1)$ OLS confidence interval: $c.i.(\widehat{\beta}_1)_{OLS} = [0.334; 0.839]$. The

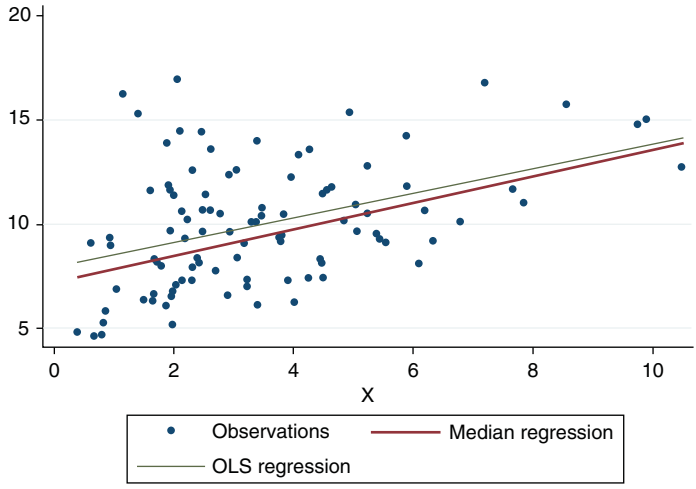

Figure 3.20 OLS and median estimated regression of the model with simulated data. The OLS line is attracted by the larger outliers in Y and moves upward. Sample size n = *100. The two lines have similar slope and differing intercepts, however the QR estimates do not statistically differ from the OLS results.*

OLS intercept is $\widehat{\beta}_0 = 7.943$ and is greater than the intercept of the median regression, $\widehat{\beta}_0(0.50) = 7.240$. The later falls still inside the $\pm 2\widehat{se}(\widehat{\beta}_0)$ OLS confidence interval, $c.i.(\widehat{\beta}_0)_{OLS} = [6.923; 8.964]$, although it is quite close to the lower bound, and the difference is statistically irrelevant.

The median regression is only one of the QRs that is possible to compute. Therefore in the above examples, besides the median, other quantiles are selected: the 10th, the 25th (first quartile), the 75th (third quartile), and finally the 90th quantile. A finer subdivision, involving a greater selection of quantiles, can be considered and easily computed, as detailed in Chapter 2.

Figure 3.21 and Table 3.3 present the QR results for the wage equation. In the graph the fitted lines show that the different QR estimates produce parallel lines, which differ from one another only by the intercept. In Table 3.3 the estimated slope at the different quantiles goes from a minimum of $\widehat{\beta}_1(.25) = 0.044$ to a maximum of $\widehat{\beta}_1(.50) = \widehat{\beta}_1(.75) = \widehat{\beta}_1(.90) = 0.091$, all of them positive, thus confirming the presence of an upward slope at all conditional quantiles. The implications of changing estimates across quantiles will be discussed shortly. First the statistical relevance of the QR estimates has to be assessed. Focusing on the slope, the computed Student-*t* values are: $t(\widehat{\beta}_1(.10)) = 2.03$, $t(\widehat{\beta}_1(.25)) = 2.67$, $t(\widehat{\beta}_1(.75)) = 3.43$, $t(\widehat{\beta}_1(.90)) = 2.38$. All these values allow to reject the null $H_0 : \beta_1(\theta) = 0$, and all these coefficients are statistically different from zero. Thus, the proportionality between log of wage and education is positive and statistically relevant at all quantiles. In addition, it is greater at and above the median; that is education grants higher

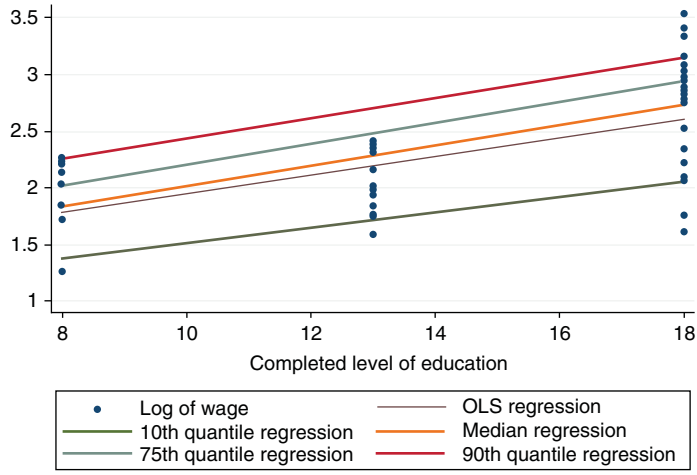

Figure 3.21 QRs and OLS estimates of the wage equation. The estimated QRs are parallel to one another, while the OLS estimated line has a slightly different slope.

returns at and above the median wage. The intercepts are statistically significant as well at all but the first quantile.

The next question is whether the QR estimates are statistically different from the OLS coefficient, that is if they lie outside the OLS confidence interval. All the QR estimates fall within the $\pm 2\widehat{se}(\widehat{\beta}_1)$ OLS confidence interval, which for the slope is given by $c.i.(\widehat{\beta}_1)_{OLS} = [0.0481; 0.113]$, with the sole exception of $\widehat{\beta}_1(.25)$. Figure 3.21, comparing OLS and QR estimates, shows an OLS slope slightly different from the parallel lines of the QRs. However such difference is statistically irrelevant. The estimated QR intercepts are all within the $\pm 2\widehat{se}(\widehat{\beta}_0)$ OLS confidence interval, given by $c.i.(\widehat{\beta}_0)_{OLS} = [0.6572; 1.642]$. It is worth noting that the OLS estimates assume an intermediate value between the above-the-median and the below-the-median QR results.

Besides the comparison between QRs and OLS results, it is also meaningful to compare a QR with all the other quantiles. Indeed, while in Figure 3.21 the QRs look like parallel lines, the estimates reported in Table 3.3 for the slope coefficient are not exactly equal across quantiles. It is important to assess if this inequality is statistically relevant, or if it is only due to the small size of the sample. A measure of the statistical significance of the discrepancy of a coefficient estimated at different quantiles is provided by the interquantile difference. For any two quantiles θ_1 and θ_2, the interquantile difference provides an estimate of:

$$Y(\theta_2) - Y(\theta_1) = [\beta_0(\theta_2) - \beta_0(\theta_1)] + [\beta_1(\theta_2) - \beta_1(\theta_1)]X = \gamma + \delta X$$

where the values of the dependent and independent variables at the different quantiles $Y(\theta_1)$ and $Y(\theta_2)$ are provided by bootstrap. The coefficient γ, given by $[\beta_0(\theta_2) - \beta_0(\theta_1)]$, measures the difference in the intercept from one quantile to the other, while

$\delta = [\beta_1(\theta_2) - \beta_1(\theta_1)]$ computes the difference between the slope coefficient at the two selected quantiles. In the interpretation of γ and δ, a change in the intercept is not really a surprise, since by changing quantile the estimated line moves along the vertical axis: when the explanatory variable is equal to zero the intercept reports the selected quantile of Y, as discussed in Chapter 1. The comparison of the estimated γ in the tails is quite informative since it can signal the presence of skewness in the conditional distributions. However the line can shift in a parallel way, so that the fitted lines at different quantiles have the same slope, and $\delta = 0$, or it can shift by changing both intercept and slope, thus breaking the parallelism. Chapter 5 will discuss the implications of a slope changing across quantiles. Here the focus relies on verifying the presence of parallel shifts of the fitted line since, when this is the case, the data are i.i.d. Conversely, when $\delta \neq 0$ and the slope at different quantiles does change, the i.i.d. assumption is groundless. The statistical relevance of the difference between the estimated slopes at two different quantiles, as estimated by $\widehat{\delta}$, can be checked with the usual inference tools, standard errors and Student-t statistics of $\widehat{\delta}$. When the null $H_0 : \delta = 0$ is rejected, the difference between the estimated slopes is statistically different from zero, the two QRs have statistically different slopes, the fitted lines are not parallel and the data are not i.i.d. The interquantile results for the wage equation are reported in Table 3.4. The table shows that the differences in the slope coefficients as computed at adjacent quantiles are not statistically different from zero, and even the comparison of two distant quantiles, the first and the last one, yields a non statistically significant difference in the estimated slopes. The conclusion is that the quantile regressions of wage as a function of education, in the small subset of $n = 49$ observations selected, when computed at different quantiles yield parallel fitted lines and the i.i.d. assumption is valid.

Moving to the other example, Figure 3.22 and Table 3.5 report the results for the simulated data. The graph shows estimated QRs which tend to converge toward the end of the sample, indeed in the table the estimated slopes decrease across quantiles while the OLS slope assumes an intermediate value. The table shows that the estimates are significantly different from zero and that all the QRs slopes fall within the $\pm 2\widehat{se}(\widehat{\beta}_1)$ OLS confidence interval, given by $c.i.(\widehat{\beta}_1)_{OLS} = [0.334; 0.839]$. Thus, the changes in slope are not statistically relevant. This is not the case for the intercepts, since only at the median the intercept falls within the $\pm 2\widehat{se}(\widehat{\beta}_0)$ OLS confidence interval, given by $c.i.(\widehat{\beta}_0)_{OLS} = [6.923; 8.964]$. Table 3.6, which reports the

Table 3.4 Interquantile differences in the estimated coefficients, wage equation.

	$0.25 - 0.10$	$0.50 - 0.25$	$0.75 - 0.50$	$0.90 - 0.75$	$0.90 - 0.10$
$\beta_1(\theta_2) - \beta_1(\theta_1)$	−0.023	0.046	−0.0001	−0.001	0.023
	(0.035)*	(0.027)*	(0.022)*	(0.026)*	(0.038)*
$\beta_0(\theta_2) - \beta_0(\theta_1)$	0.529	−0.251	0.193	0.205	0.676
	(0.486)*	(0.357)*	(0.287)*	(0.325)*	(0.505)*

Standard errors are given in parentheses, sample size $n = 49$. The asterisks indicate those coefficients having low Student-t statistics.

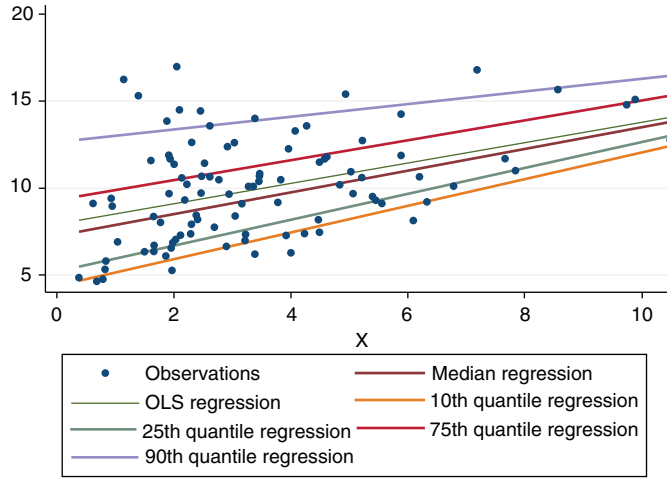

Figure 3.22 QRs and OLS estimates, simulated data. The estimated lines tend to converge at the end of the sample. The distance between the 75th and the 90th QRs is larger than the distance between the 10th and the 25th estimated lines, as measured by γ, thus signaling skewed distributions.

Table 3.5 OLS and QR estimates, simulated data.

	Estimator					
	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	$\theta = 0.90$	OLS
x_i	0.761	0.734	0.629	0.573	0.362	0.587
	(0.105)	(0.133)	(0.175)	(0.203)	(0.350)*	(0.127)
constant	4.355	5.230	7.240	9.366	12.597	7.943
	(0.465)	(0.514)	(0.707)	(0.862)	(1.591)	(0.514)

Standard errors are given in parentheses, sample size $n = 100$. The asterisk indicates a coefficient having low Student-t statistics.

Table 3.6 Interquantile differences in the estimated coefficients, simulated data.

	0.25−0.10	0.50−0.25	0.75−0.50	0.90−0.75	0.90−0.10
$\beta_1(\theta_2) - \beta_1(\theta_1)$	−0.027	−0.104	−0.056	−0.211	−0.398
	(0.091)∗	(0.150)∗	(0.203)∗	(0.269)∗	(0.380)∗
$\beta_0(\theta_2) - \beta_0(\theta_1)$	0.875	2.009	2.126	3.231	8.243
	(0.467)	(0.733)	(1.061)	(1.442)	(1.755)

Standard errors are given in parentheses, sample size $n = 100$. The asterisks indicate those coefficients having low Student-t statistics.

interquantile estimates, confirms that the QRs can be considered parallel lines: the interquantile slopes are not significantly different from zero as they should be, since by construction the data are i.i.d. However in this model the estimated γ are all significantly different from zero. The great dispersion in Y causes great shifts in the estimated QRs. As mentioned, the comparison of γ as estimated in the tails helps to spot skewness in the conditional density. While in Table 3.4 the interquantile differences for the intercept are not statistically relevant, in Table 3.6 they are statistically different from 0. Moreover, the comparison of γ as estimated at (0.25–0.10) and at (0.90–0.75) clearly shows the presence of skewness, since the estimated value of γ at the lower tail (0.25–0.10) is much smaller than its estimate at the upper tail (0.90–0.75). The simulated data are indeed characterized by errors generated by a χ^2 with 5 degrees of freedom.

3.3 Wald, Lagrange multiplier, and likelihood ratio tests

This section further analyses the small dataset on wages. While the previous section considers education as the sole explanation of log of wage, now other explanatory variables are included in the equation in order to gain more insight into the factors driving earnings.

Economic theory, since Mincer (1974), relates wage to age, as a proxy of experience, quite unfortunately but realistically to gender, where women earn less, and to education. Of course other variables concur to define wage and the model can become quite sophisticated. For instance the dataset analyzed here includes other variables, such as hours of independent work and part time work that can all be considered as valid explanatory variables in a wage equation. In the small sample selected here, age assumes values from 25 to 60; gender is a dummy assuming unit value for 13 women out of the 49 interviewed persons; number of hours of autonomous work, 'independent' in the equation, assumes values from 3 to 72 and is very spread out, with a mode of 40 h having frequency 7, as can be seen in Figure 3.23; part time job, 'partime' in the equation, is a dummy taking unit value in 15 out of 49 observations. The equation becomes:

$$\begin{aligned} lwage_i = \beta_0 + \beta_1 education_i + \beta_2 age_i + \beta_3 gender_i \\ + \beta_4 partime_i + \beta_5 independent_i + e_i; \end{aligned}$$

and the results of quantile and OLS regressions are reported in Table 3.7. A quick comparison of these results with the estimates in Table 3.3, where only education would explain earnings, shows that the education coefficients are now lower, no matter what estimator is considered. In Table 3.7, however, most of the estimated coefficients are not statistically different from zero, particularly at the lower quantiles, and some of the new explanatory variables can be safely excluded. The exclusion of a single explanatory variable can be decided on the basis of the Student-t test.

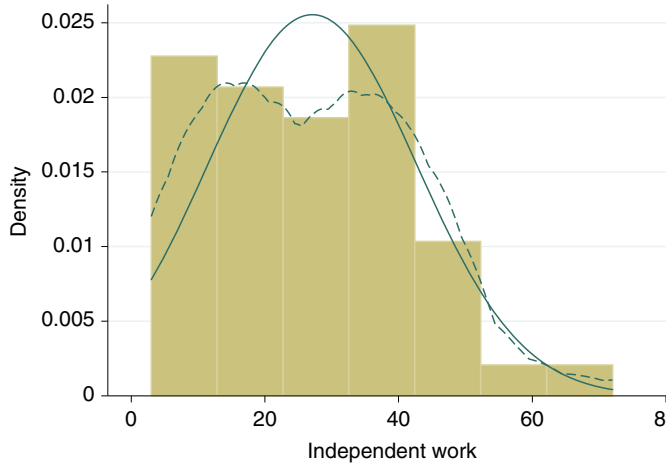

Figure 3.23 Histogram of the explanatory variable number of hours of independent work.

Table 3.7 Estimated coefficients, wage equation.

	Estimator					
	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	$\theta = 0.90$	OLS
education	0.027	0.026	0.053	0.071	0.091	0.068
	(0.016)∗	(0.018)∗	(0.023)	(0.028)	(0.028)∗	(0.015)
age	0.013	0.013	0.027	0.027	0.026	0.023
	(0.011)∗	(0.011)∗	(0.009)	(0.008)	(0.010)	(0.006)
gender	−0.136	−0.115	−0.092	0.189	−0.050	−0.083
	(0.137)∗	(0.184)∗	(0.199)∗	(0.176)∗	(0.148)∗	(0.125)∗
part time	−0.234	0.232	0.209	0.144	0.214	0.197
	(0.194)∗	(0.178)∗	(0.184)∗	(0.184)∗	(0.226)∗	(0.120)∗
independent	−0.012	−0.006	0.0009	0.002	0.007	−0.001
	(0.004)∗	(0.006)∗	(0.006)∗	(0.006)∗	(0.005)∗	(0.004)∗
constant	1.449	1.409	0.492	0.059	0.209	0.472
	(0.690)∗	(0.596)	(0.620)	(0.777)∗	(0.953)∗	(0.389)

Standard errors are given in parentheses, sample size $n = 49$. The asterisk indicate those coefficients having low Student–t statistics.

It is possible to sequentially exclude one variable at the time from the above equation, but which one is the variable to exclude first? Having many QRs does not make this task easy. The only way is to test, and eventually exclude, jointly and not sequentially the variables of the model. To jointly exclude two or more variables a different test is needed. The Wald (W), Lagrange multiplier (LM) and likelihood ratio (LR) tests, which are asymptotically equivalent and asymptotically distributed as a χ^2, allow to verify the exclusion of more than one coefficient at a time. The degrees of freedom of the χ^2 are equal to the number of coefficients under test.

For instance, gender and hours of autonomous work are not statistically significant in any of the equations reported in Table 3.7, and it is relevant to test the null $H_0 : \beta_3 = \beta_5 = 0$. This hypothesis involves the introduction of constraints which set to zero two coefficients at a time, thus implying the exclusion of the corresponding explanatory variables.

The test functions for the LR test is defined as:

$$LR = 2\omega^{-1}(\widetilde{V}(\theta) - \widehat{V}(\theta)). \tag{3.4}$$

The LR test relies on the comparison of the estimated objective function of the equation excluding the explanatory variables under test, the so-called restricted model $\widetilde{V}(\theta)$, and the estimated objective function of the equation including them, the unrestricted model $\widehat{V}(\theta)$ (Koenker and Bassett 1982b). If the variables are incorrectly excluded, the constrained model will have a larger objective function, the difference $[\widetilde{V}(\theta) - \widehat{V}(\theta)]$ will be statistically relevant and the null is rejected. Conversely, when the excluded variables are irrelevant, the two objective functions are close to one another, the restrictions are valid and the null is not rejected. If the two objective functions are close enough the restrictions are valid. The scale parameter ω is defined as $\omega^2(\theta) = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2}$. At the median, to test the null $H_0 : \beta_3 = \beta_5 = 0$, the two objective functions, respectively, for unrestricted and restricted model, are

$$\begin{aligned}
\widehat{V}(\theta) = \sum\nolimits_i |&lwage_i - \beta_0 - \beta_1 education_i - \beta_2 age_i - \beta_3 gender_i \\
&- \beta_4 partime_i - \beta_5 independent_i| \,, \\
\widetilde{V}(\theta) = \sum\nolimits_i |&lwage_i - \beta_0 - \beta_1 education_i - \beta_2 age_i - \beta_4 partime_i
\end{aligned}$$

The residuals of these two QR estimated models are shown in Figure 3.24. The statistical relevance of the difference between these two graphs can be ascertained by the comparison of the estimated objective functions, which assume, respectively, the values of $\widehat{V}(\theta) = 13.39$ and $\widetilde{V}(\theta) = 13.52$. For the constrained model $\omega(\theta) = 0.50/0.93 = 0.54$ and $LR = 3.71(13.52 - 13.39) = 0.46$, to be compared with the critical value 5.99 of a χ^2 with 2 degrees of freedom at the 5% significance level. Thus the null is not rejected, restricted and unrestricted models yield similar results, and the variables gender and autonomous work can be safely dropped.

However, the part time coefficient is marked with an asterisk in each column of Table 3.7, and can be possibly excluded as well. To test the exclusion of three variables together, that is to check the relevance of gender, part time and hours of autonomous work and verifying if all of them are non-significantly different from zero in the estimated equation, the hypothesis under test is $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. At the median this is tested by comparing the previous unconstrained objective function $\widehat{V}(\theta)$ with the following constrained objective function

$$\widetilde{V}(\theta) = \sum\nolimits_i |lwage_i - \beta_0 - \beta_1 education_i - \beta_2 age_i| \,.$$

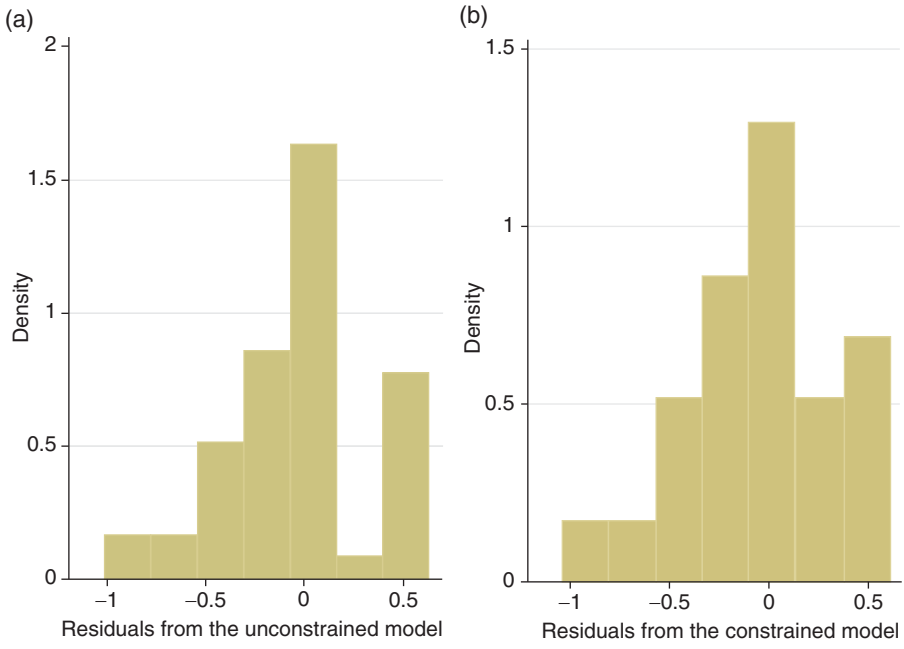

Figure 3.24 Residuals from the (a) unconstrained and (b) constrained model, to test the null $H_0 : \beta_3 = \beta_5 = 0$.

Figure 3.25 reports the histograms of the residuals from the unrestricted and restricted models to test the exclusion of these three explanatory variables. The estimated objective functions are $\widehat{V}(\theta) = 13.39$ and $\widetilde{V}(\theta) = 14.22$. For the constrained model $\omega(\theta) = 0.50/0.94 = 0.53$ and $LR = 3.76(14.22 - 13.39) = 3.12$, which does not reject the null when compared with the critical value of a χ^2 with 3 degrees of freedom, equal to 7.81 at the 5% level.

Another test to verify the hypothesis on more than one coefficient at a time is the LM test. It considers the gradient $\mathbf{g}$ of the model excluding the variables under test. The test function is

$$LM = \mathbf{g}^T [\mathbf{D}_{22}]^{-1} \mathbf{g}, \tag{3.5}$$

and for small values of the estimated test function the variables under test can be safely excluded from the equation (Koenker and Basset 1982b). As shown in Chapter 1, the objective function of the median regression is given by $\sum_i \rho_\theta(e_i) = 0.5 \sum_i |e_i|$. The gradient at the median, that is the first order condition with respect to the coefficient of the generic explanatory variable X at the median, is given by $\sum_i \rho'_\theta(e_i) = \sum_i \psi(e_i) x_i = \sum_i sgn(e_i) x_i$. That is, the gradient is a function of the sign of the errors, of their position above or below the QR line. Their numerical value is not relevant, but their position with respect to the fitted line and thus their sign

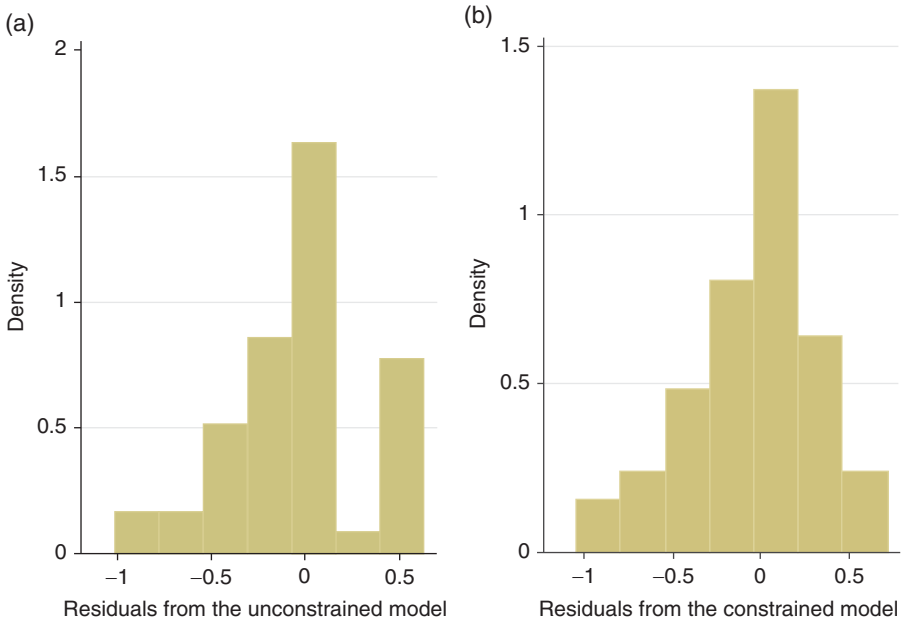

Figure 3.25 Residuals from the (a) unconstrained and (b) constrained model, to test the null H_0: $\beta_3 = \beta_4 = \beta_5 = 0$.

is crucial. The $\mathbf{D}$ matrix, a quadratic form comprising all the explanatory variables $\mathbf{D} = \mathbf{X}^T\mathbf{X}$, is partitioned in blocks $\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{bmatrix}$, the vector $\mathbf{g}$ together with the submatrix $\mathbf{D}_{22}$ select the variables under test, and its inverse $\mathbf{D}^{22}$ is given by $\mathbf{D}^{22} = [\mathbf{D}_{22} - \mathbf{D}_{21}\mathbf{D}_{11}\mathbf{D}_{12}]^{-1}$.

The LM test can also be implemented by estimating an auxiliary regression. The residuals of the constrained model become the dependent variable of an additional regression having as explanatory variables those regressors excluded from the model (Weiss 1990). The term nR^2 is asymptotically χ^2 with degrees of freedom equal to the number of variables under test. The auxiliary regression checks if the excluded regressors have any explanatory content that would be lost once they were eliminated from the main equation. If the variables are erroneously excluded, they will explain at least part of the residuals from the main equation, the auxiliary regression will have a large nR^2 and the null on the validity of the constraints will be rejected. Coversely, if the regressors under test are superfluous, in the auxiliary equation the nR^2 term is small and the null is not rejected. Under the null $H_0 : \beta_3 = \beta_5 = 0$, the residuals of the constrained model estimated at the median:

$$\tilde{e}_i(\theta) = lwage_i - \beta_0 - \beta_1 education_i - \beta_2 age_i - \beta_4 partime_i$$

become the dependent variable in the auxiliary equation:

$$\widetilde{e}_i(\theta) = \alpha_0 + \alpha_1 gender_i + \alpha_2 independent_i + \eta_i$$

where $nR^2 = 0.87$ is compared with the tabulated value of $\chi^2_{(2)} = 5.99$ at the 5% level. Thus the two variables do not have any explanatory content and can be safely eliminated. In order to test the null $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$, the residuals from the restricted model:

$$\widetilde{e}_i(\theta) = lwage_i - \beta_0 - \beta_1 education_i - \beta_2 age_i$$

become the dependent variable in the auxiliary equation:

$$\widetilde{e}_i(\theta) = \alpha_0 + \alpha_1 gender_i + \alpha_2 partime_i + \alpha_3 independent_i + \eta_i$$

where $nR^2 = 3.54$ is lower than the tabulated value of $\chi^2_{(3)} = 7.81$ at the 5% level. The null is not rejected and the variable $partime_i$ can be safely excluded as well.

Finally the Wald test, denoted by W, looks at the estimates of the model including the variables under test (Koenker and Basset 1982b). If their quadratic function is close to zero the variables under test can be safely excluded. The test function is:

$$W = n\omega^{-2}\widehat{\boldsymbol{\beta}}(\theta)^\top \left[\mathbf{D}^{22}\right]^{-1} \widehat{\boldsymbol{\beta}}(\theta). \tag{3.6}$$

To test the null $H_0 : \beta_3 = \beta_5 = 0$, the vector $\widehat{\boldsymbol{\beta}}(\theta)$ of the estimated coefficients under test, is given by: $\widehat{\boldsymbol{\beta}}(\theta) = \begin{bmatrix}\widehat{\beta}_3 \\ \widehat{\beta}_5\end{bmatrix} = \begin{bmatrix} -0.0927 \\ 0.0009 \end{bmatrix}$, the matrices $\mathbf{D}_{22}$ and $\left[\mathbf{D}^{22}\right]^{-1}$ are:

$$\begin{aligned}\mathbf{D}_{22} &= \begin{bmatrix} \sum_i gender_i^2 & \sum_i gender_i independent_i \\ \sum_i independent_i gender_i & \sum_i independent_i^2 \end{bmatrix} \\ &= \begin{bmatrix} 88 & 1705 \\ 1705 & 47694 \end{bmatrix}, \\ n\left[\mathbf{D}^{22}\right]^{-1} &= \begin{bmatrix} 9.44 & 27.40 \\ 27.40 & 10088.8 \end{bmatrix}.\end{aligned}$$

The scale for the unrestricted model is estimated as $\widehat{\omega}^2 = 0.25/0.607 = 0.41$ and the test function $W = 0.085/0.41 = 0.20$ to be compared with $\chi^2_{(2)} = 5.99$ at the 5% level. The null is not rejected and the two variables can be excluded from the model. To test the hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$, the estimated vector of the coefficients under test is $\widehat{\boldsymbol{\beta}}(\theta) = \begin{bmatrix}\widehat{\beta}_3 \\ \widehat{\beta}_4 \\ \widehat{\beta}_5\end{bmatrix} = \begin{bmatrix} -0.0927 \\ 0.209 \\ 0.0009 \end{bmatrix}$. The matrices $\mathbf{D}_{22}$ and $\left[\mathbf{D}^{22}\right]^{-1}$ are:

$$\mathbf{D}_{22} =$$

$$= \begin{bmatrix} \sum_i gender_i^2 & \sum_i gender_i partime_i & \sum_i gender_i independent_i \\ \sum_i partime_i gender_i & \sum_i partime_i^2 & \sum_i partime_i independent_i \\ \sum_i independent_i gender_i & \sum_i independent_i partime_i & \sum_i independent_i^2 \end{bmatrix}$$

$$= \begin{bmatrix} 88 & 20 & 1705 \\ 20 & 15 & 383 \\ 1705 & 383 & 47694 \end{bmatrix},$$

$$n\left[\mathbf{D}^{22}\right]^{-1} = \begin{bmatrix} 9.538 & 0.989 & 25.28 \\ 0.989 & 10.33 & -22.1 \\ 25.28 & -22.1 & 10136.2 \end{bmatrix}.$$

The estimated Wald test is $W = 0.493/0.41 = 1.20$. At the 5% level, the null is not rejected when compared with $\chi^2_{(3)} = 7.81$.

All three tests agree to eliminate the *gender*, *independent* and *partime* variables.

While the LR test computes both restricted and unrestricted models, the LM test considers only the restricted version and the Wald test looks at the unrestricted equation. As mentioned, the three tests are asymptotically equivalent and are all asymptotically distributed as χ^2 with degrees of freedom equal to the number of constrained coefficients.

The tests here implemented are nested, that is the hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ embeds the previous one, $H_0 : \beta_3 = \beta_5 = 0$. This does not necessarily need to be the case, and the tests for exclusion restrictions may very well be non-nested.

For a further check, it is possible to test the exclusion of one additional variable, in turn age or education. By testing $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ the estimated test functions are: $LR = 2.51(17.42 - 13.39) = 10.11$; $W = 14.82$; and $LM = 12.82$. For the test of the null $H_0 : \beta_1 = \beta_3 = \beta_4 = \beta_5 = 0$ the computed tests are: $LR = 4.05(16.88 - 13.39) = 14.13$, and $LM = 17.32$. In both the cases the null is rejected when compared with the critical value 9.48 of a χ^2_4 at the 5% significance level.[3] Figure 3.26 and Figure 3.27 compare the residuals of the constrained and unconstrained models for these last two tests.

Thus, the final definition of the wage equation is $lwage_i = \beta_0 + \beta_1 education_i + \beta_2 age_i + e_i$. The estimates are reported in Table 3.8. At the 10th quantile, all the regression coefficients are not statistically relevant. The *age* coefficient is not significantly different from 0 at the 25th quantile, signaling that at low *wages*, experience is not really a relevant explanatory variable. However, at and above the median, both *education* and *age* are statistically relevant. In particular, the *education* coefficient grows from 6.4 to 8.5% in going from the median to the upper quantile, which can be compared with the OLS estimate, that assumes the intermediate value of 7.0%. The QR results show that while at the lower wages education is not relevant, at higher wages returns for education increase. The QR estimates fall

[3] The sole exception is the Wald test for the null $H_0 : \beta_1 = \beta_3 = \beta_4 = \beta_5 = 0$ which yields $W = 8.37$.

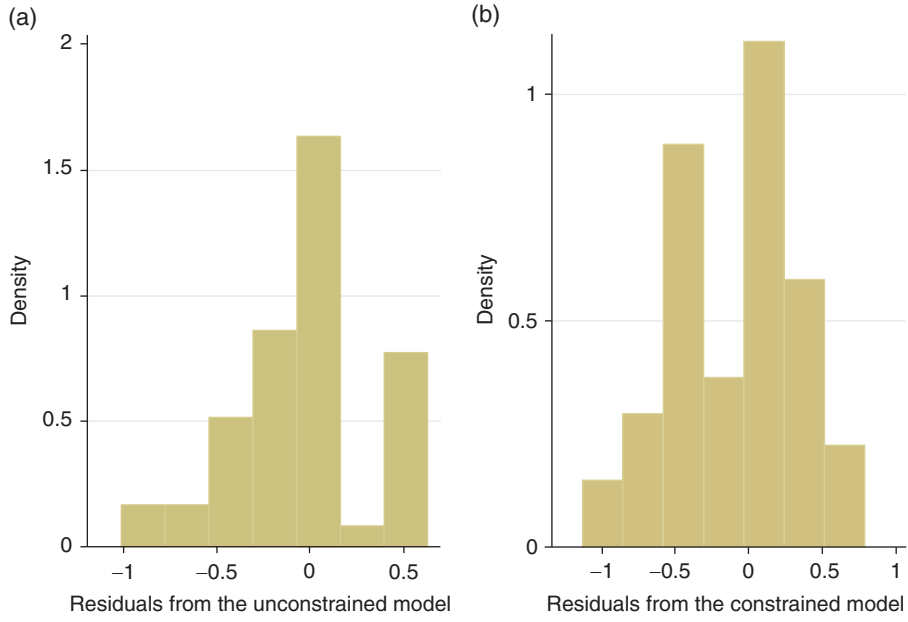

Figure 3.26 Residuals from the (a) unconstrained and (b) constrained model, to test the null H_0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

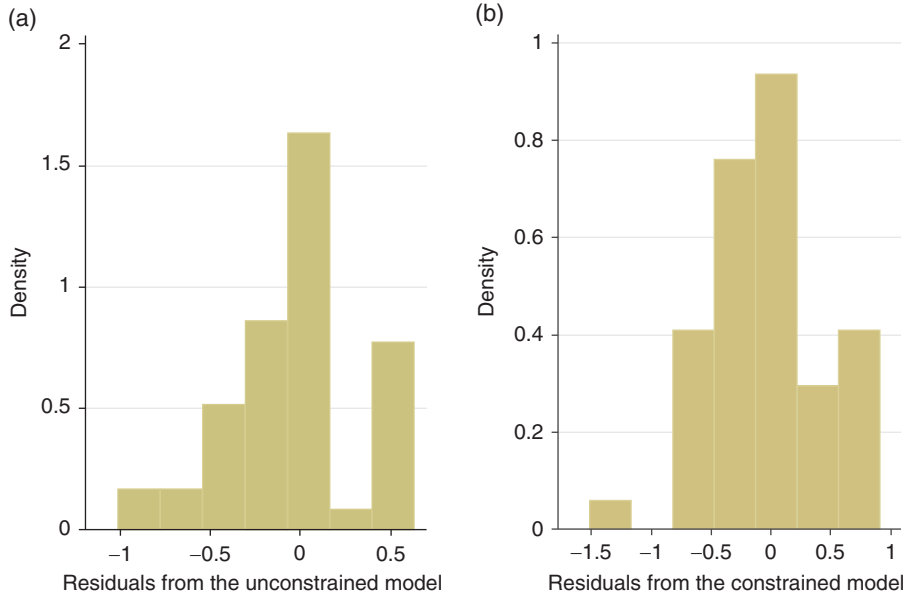

Figure 3.27 Residuals from the (a) unconstrained and (b) constrained model, to test the null H_0: $\beta_1 = \beta_3 = \beta_4 = \beta_5 = 0$.

Table 3.8 Estimated coefficients, wage equation.

	Estimator					
	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	$\theta = 0.90$	OLS
education	0.051	0.056	0.064	0.083	0.085	0.070
	(0.051)∗	(0.021)	(0.019)	(0.024)	(0.023)	(0.014)
age	0.019	0.011	0.027	0.020	0.028	0.022
	(0.020)∗	(0.010)∗	(0.008)	(0.008)	(0.008)	(0.006)
constant	0.326	0.693	0.326	0.490	0.404	0.380
	(0.398)∗	(0.406)∗	(0.394)∗	(0.524)∗	(0.573)∗	(0.296)∗

Standard errors are given in parentheses, sample size $n = 49$. The asterisks indicate those coefficients having low Student-t statistics.

Table 3.9 Interquantile differences in the estimated coefficients, wage equation.

	0.25−0.10	0.50−0.25	0.75−0.50	0.90−0.75	0.90−0.10
$\beta_1(\theta_2) - \beta_1(\theta_1)$	−0.004	0.008	0.018	0.002	0.033
	(0.029)∗	(0.017)∗	(0.017)∗	(0.020)∗	(0.031)∗
$\beta_2(\theta_2) - \beta_2(\theta_1)$	−0.007	0.015	−0.006	0.007	0.008
	(0.007)∗	(0.007)	(0.008)∗	(0.012)∗	(0.009)∗

Standard errors are given in parentheses, sample size $n = 49$. The asterisks indicate those coefficients having low Student-t statistics.

inside the $\pm 2\sigma(\beta)$ OLS confidence interval. Table 3.9 provides the estimates of the interquantile differences, with estimated coefficients non-significantly different from zero for both the education and the age coefficient, respectively, $[\beta_1(\theta_2) - \beta_1(\theta_1)]$ and $[\beta_2(\theta_2) - \beta_2(\theta_1)]$. In the case of more explanatory variables the analysis of the interquantile differences is a quick and easy tool to control the parallelism of the fitted planes, that is to control the assumption of i.i.d. conditional distributions.

3.4 Summary of key points

- The QR estimator is asymptotically normal as shown by analyzing the empirical distributions of the slope coefficient in a simple linear regression model. A simulation study is implemented considering: (i) i.i.d. data, with both symmetric and skewed distributions; (ii) i.ni.d. data; (iii) dependent data.
- In the case of normality the OLS estimator is more efficient, but when the distributions are non-normal the precision of the QR estimator improves upon OLS.
- Inference is implemented for each coefficient with the Student-t test. Confidence intervals allow to compare QRs and OLS estimates in order to assess the equality or inequality of the results.

- The comparison between different QR estimates is implemented by looking at the interquantile differences. This is a relevant issue since QRs may differ only in the intercept or in both intercept and slope and the interquantile differences allow to state which one is the case. In the former the conditional distribution of the dependent variable is i.i.d., conversely in the latter the i.i.d. assumption is unrealistic.
- When the interquantile difference as computed at the lower tail diverges from its analog computed at the upper tail, the conditional distribution is skewed.
- In regressions with many explanatory variables, inference on more than one coefficient at a time is implemented by Wald, LM and LR tests for QRs.

References

Epanechnikov V 1969 Nonparametric estimation of a multidimensional probability density. *Theory of Probability and its Applications* **14**(1), 153–158.

Koenker R 2005 *Quantile Regression*. Cambridge University Press.

Koenker R and Basset G 1978 Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker R and Basset G 1982a Robust tests for heteroskedasticity based on regression quantiles. *Econometrica* **50**, 43–61.

Koenker R and Basset G 1982b Tests for linear hypotheses and L_1 estimation. *Econometrica* **50**, 1577–1583.

Mincer J 1974 *Schooling, Experience and Earnings*. Columbia University Press.

Weiss A 1990 Least absolute error estimation in the presence of serial correlation. *Journal of Econometrics* **44**, 127–159.

4

Additional tools for the interpretation and evaluation of the quantile regression model

Introduction

To appreciate the meaningful potentialities of quantile regression (QR), it is necessary to have a greater understanding of the interpretations and the evaluation tools described in the previous chapters. In real dataset applications, the analysis requires the adoption of several proper choices, starting from the preparation of the input data through the description and interpretation of the results until the model validation.

This chapter deals with some typical issues arising from a real data analysis, highlighting the capability of QR and its differences compared with other methods. In particular, the effect of variable centring and scaling on the interpretation of the results is explored, both from a descriptive and from an inferential point of view. The interpretation of QR results can be refined and enhanced by the estimation of the conditional density of the response variable. Finally, the capabilities of bootstrap methods are also explored to estimate standard errors in QR. To appreciate differences and peculiarities of bootstrap methods, homogeneous and heterogeneous variance regression models are considered.

In the following, four real datasets are used that best illustrate the interpretation and the potentiality of QR results: the *IlSole24ore* dataset, the *obese* dataset, the *degree* dataset and the *satisfaction* dataset.

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

4.1 Data pre-processing

In real data analysis, transformation of the response and/or the explanatory variables can often be advisable to handle skewness in the data, missing data, and outliers, to introduce nonlinearity and facilitating the interpretation of the model. QR estimators exhibit several properties when the response and/or the explanatory variables are subject to mathematical transformations. Some data transformations are also typical of least squares estimators. However, caution is often required when interpreting the results because the same rules applied to the original data cannot be applied when transforming the results back to the original scale because this might lead to biased results.

In this section the main properties of QR estimators when the response and/or the explanatory variables are subject to mathematical transformations are described.

4.1.1 Explanatory variable transformations

Classical transformation of the explanatory variables is their centring. This allows the intercept coefficient to be interpreted in a reasonable manner. In a standard linear regression, the intercept enables measurement of the dependent variable value, which is derived form setting all the explanatory variables to zero. In most practical cases, the interpretation of the intercept is ambiguous because it is difficult to assume a value equal to zero for the independent variables.

With these factors in mind, let us consider the *IlSole24ore* dataset and refer to the simplest linear model. The *rate of suicide* (hereafter referred to as *suicide*) is considered as the dependent variable. The *mean per capita income* (hereafter referred to as *income*) is used as the explanatory variable.

Table 4.1 shows the ordinary least squares (OLS) and QR estimates obtained from the original data (first and second rows) and centring the *income* variable (third and fourth rows).

Considering the original data, where the standard linear regression results show that *suicide* increases, on average, by 0.38 points for every one unit increase in *income*, the intercept is equal to 3.11.

Table 4.1 OLS and QR coefficients for the original and the centred *income* variable. Centring the explanatory variable permits the intercept interpretation without altering the magnitude and the direction of the coefficients.

Data		OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
Original	(Intercept)	3.11	1.32	0.57	1.51	5.45	−1.26
	income	0.38	0.22	0.31	0.38	0.40	0.77
Centred	(Intercept)	16.65	9.28	11.45	15.10	19.79	26.08
	income	0.38	0.22	0.31	0.38	0.40	0.77

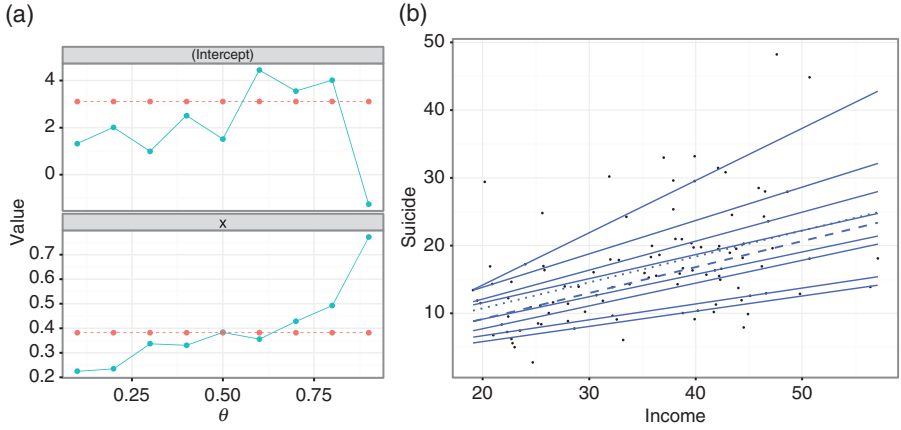

Figure 4.1 QR coefficient plots (a) corresponding to quantiles: $\theta = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ *and scatter plot of* income *and* suicide *with the estimated QR and OLS lines (b). In the coefficient plot, quantiles are represented on the horizontal axis. In the scatter plot, the dashed line represents the median line, the solid lines the other quantile lines and the dotted line the OLS regression line. Moving from lower to upper quantiles, the contribution of* income *to* suicide *increases.*

For a selected number of quantiles, the effect of *income* increases, moving from the lower part to the upper part of the *suicide* distribution data [Figure 4.1(a)], sign of a location-scale model namely of changes both in the central tendency and in the variability of the dependent variable. If a one-unit increase in the *income* variable in the central location (both in mean and in the median) leads to a 0.38 increase in the *rate of suicide*, the rate decreases to 0.22 at the 10th quantile and increases to 0.77 in the 90th quantile. In practical terms, the QR results suggest that people living in provinces where the *rate of suicide* is very high are more likely to be prone to suicide if their income is higher than those people living in provinces where the *rate of suicide* is very low.

In Figure 4.1(b), the regression lines corresponding to the selected quantiles (dashed line for the median and solid lines for the other quantiles) together with the OLS regression line (dotted line) are superimposed on the original scatter plot. The slopes increase, moving from the lower to the higher quantiles. The higher slopes corresponding to the highest quantiles reveal greater variability in the conditional distribution for provinces with high *rates of suicide*.

The intercept interpretation is ambiguous because it is difficult to assume a value equal to zero for the *income* variable. To overcome this drawback, the analysis can be performed on the explanatory variable centred data. Centring the explanatory variable does not change the magnitude and the direction of the coefficients, but it allows interpretation of the intercept coefficient values: in the case of a standard linear regression, it measures how much the average *rate of suicide* increases for a

province with an average *per capita income*, whereas for each quantile, it measures how much the given quantile of the *rate of suicide* increases for a province with an average *per capita income*.

Table 4.1 compares, for a selected number of quantiles, the OLS and the QR coefficients, obtained for the original and the centred *income* variable.

4.1.2 Dependent variable transformations

One of the main advantages of QR estimators is their behavior with respect to monotone transformations of the response variable. This behavior, named *equivariance*, refers to the ability to use the same interpretation rules when the data or the model are subjected to a transformation. Some authors have proposed that the equivariance property can be exploited to speed up the estimation process by reducing the number of simplex iterations (Buchinsky 1998).

According to the chosen transformation, the equivariance property can be distinguished in:

- scale equivariance;
- shift or regression equivariance;
- equivariance to reparametrization of design;
- equivariance to monotone transformations.

While the first three properties are also satisfied by OLS estimators, the last one is peculiar to QR, and this represents one of the strong points of the method.

In the following, equivariance properties are described, and their practical implications are highlighted. For the corresponding proofs see Koenker and Basset (1978).

Let us consider the simplest QR model with one explanatory variable and for a given quantile θ:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}. \tag{4.1}$$

The *scale equivariance* property implies that, if the dependent variable is multiplied by a positive constant c, the coefficients of the new model can be easily obtained multiplying by c the coefficients in Equation (4.1):

$$Q_\theta(c\hat{\mathbf{y}}|\mathbf{x}) = c\hat{\beta}_0(\theta) + c\hat{\beta}_1(\theta)\mathbf{x}. \tag{4.2}$$

The same property holds if the dependent variable is multiplied by a negative constant, d. In this case, the coefficients of the new model are the complement of the coefficients in (4.1) because positive and negative errors switch:

$$Q_\theta(d\hat{\mathbf{y}}|\mathbf{x}) = d\hat{\beta}_0(1-\theta) + d\hat{\beta}_1(1-\theta)\mathbf{x}. \tag{4.3}$$

For $\theta = 0.5$, the QR estimates are scale equivariant, irrespective of the sign of the constant.

The scale equivariance property can be very useful when it is necessary to modify the unit of measurement of the dependent variable (e.g., to reduce its variability or to allow comparisons with other models) without altering the interpretation of the results.

The *shift equivariance* property is also referred to as the *regression equivariance* because it denotes the effect of the dependent variable obtained as a linear combination, through the γ coefficients, of the explanatory variable. Such an effect holds when $\mathbf{y}$ is subjected to a location shift (Kuan 2007):

$$\mathbf{y}^* = \mathbf{y} + \mathbf{x}\gamma. \tag{4.4}$$

The QR estimator of $\mathbf{y}^*$ on $\mathbf{x}$ results in:

$$Q_\theta(\hat{\mathbf{y}}^*|\mathbf{x}) = \hat{\beta}_0 + \left[\hat{\beta}_1(\theta) + \gamma\right]\mathbf{x}. \tag{4.5}$$

The *equivariance to reparametrization of design* is derived from the effect of a nonsingular matrix $\mathbf{A}(p \times p)$ introduced in the model:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{XA}) = \mathbf{A}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}(\theta), \tag{4.6}$$

where $\mathbf{X}$ is the matrix of p explanatory variables.

The recourse to reparametrization is quite common in regression analysis (von Eye and Schuster 1998) when the matrix of the explanatory variables is not of full column rank. This drawback typically occurs in cases of qualitative explanatory variables. There are several ways of overcoming this deficiency by coding qualitative variables.

Let us consider the simplest QR model with *suicide* as a dependent variable and a dummy regressor represented by the difference in Celsius degrees between the hottest and the coldest month of the year (hereafter referred to as *temperature*). The two categories of the *temperature* variable are the *temperature low*, corresponding to differences less than 20°C, and the *temperature high*, corresponding to differences equal or more than 20°C.

The explanatory matrix can be expressed (Koenker and Basset 1982) as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_{low}} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_{high}} \end{bmatrix}, \tag{4.7}$$

where n_{low} and n_{high} are the number of provinces, respectively, with *low* and *high temperature*, and the *suicide* dependent variable is split into two samples $\mathbf{y} = (\mathbf{y}_{low}, \mathbf{y}_{high})$.

The QR model without the intercept is:

$$Q_\theta\left(\hat{\mathbf{y}}|\mathbf{X}\right) = \mathbf{X}\hat{\boldsymbol{\beta}}(\theta). \tag{4.8}$$

The results for a given quantile $\theta = 0.5$ are: $\hat{\beta}_{low}(0.5) = 11.89$ and $\hat{\beta}_{high}(0.5) = 17.88$.

If an intercept is considered, a column vector must be added to the $\mathbf{X}$ matrix, with all values equal to 1. The resulting $\mathbf{X}$ matrix will not be of full rank because the intercept column is the sum of the other two columns in Equation (4.7). To introduce the intercept, it is possible to reparametrize the model by transforming the $\mathbf{X}$ matrix into $\tilde{\mathbf{X}} = \mathbf{XA}$ where $\mathbf{A} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} \end{bmatrix}$.

According to the equivariance to reparametrization of the design, parameter estimates of the QR model using $\tilde{\mathbf{X}}$ can be obtained as:

$$\hat{\tilde{\boldsymbol{\beta}}}(\theta) = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}(\theta). \tag{4.9}$$

The parameter estimates in Equation (4.8) and Equation (4.9) can be compared:

$$\begin{aligned} Q_\theta\left(\hat{\mathbf{y}}|\mathbf{X}\right) &= 11.89\mathbf{x}_{low} + 17.88\mathbf{x}_{high} \\ Q_\theta\left(\hat{\mathbf{y}}|\tilde{\mathbf{X}}\right) &= 17.88 - 5.99\mathbf{x}_{low}. \end{aligned} \tag{4.10}$$

Thus, the *suicide*-conditioned estimated values, both for the *low* and the *high* temperature, are equivalent in the two models.

Finally, the *equivariance to monotone transformations* implies that if a nondecreasing function on $\Re$, $h(.)$ is applied to the dependent variable, the quantiles of the transformed $\mathbf{y}$ variable are the transformed quantiles of the original ones:

$$Q_\theta\left[\widehat{h(\mathbf{y})}|\mathbf{x}\right] = h\left[\hat{\beta}_0(\theta)\right] + h\left[\hat{\beta}_1(\theta)\right]\mathbf{x}. \tag{4.11}$$

As stated before, this property characterizes the QR estimators with respect to those of OLS. It is very important in real data applications because appropriate selection of the $h(.)$ monotone function is necessary to manage and correct different kinds of skewness.

For instance, the logarithmic transformation is a nondecreasing function that is typically applied when the dependent variable is right-skewed. Such a transformation can only be used in cases of positive values. If values equal to zero are present, it is necessary to add a constant to all the values of the variable.

Equation (4.11), in the case of a logarithmic transformation of the response variable, can be expressed as:

$$Q_\theta\left[\widehat{log(\mathbf{y})}|\mathbf{x}\right] = log\left[\hat{\beta}_0(\theta)\right] + log\left[\hat{\beta}_1(\theta)\right]\mathbf{x}. \tag{4.12}$$

The equivariance does not hold in case of OLS regression where $log\left[E(\hat{\mathbf{y}}|\mathbf{x})\right] \neq E\left[\widehat{log(\mathbf{y})}|\mathbf{x}\right]$.

An example of the advisability of using the logarithm transformation can be shown using the *IlSole24ore* dataset where the *suicide*-dependent variable is right skewed, as is evident in Figure 4.2(a). By applying a logarithmic transformation, such skewness is clearly overcome [Figure 4.2(b)].

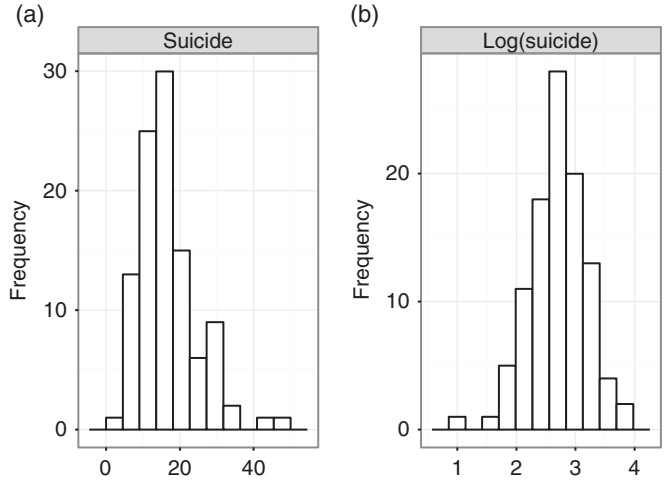

Figure 4.2 Histogram of the suicide*-dependent variable (a) and its logarithmic transformation (b). The original* suicide *variable is clearly skewed, but the asymmetry is considerably lower with the variable logarithmic transformation.*

Table 4.2 Results of White test and the Shapiro-Wilk test in an OLS regression with and without logarithmic transformation of the dependent variable. The heteroskedasticity and normality tests become significant when the dependent variable is submitted to a logarithmic transformation.

	White test	Shapiro-Wilk normality test
	(*p*-value)	(*p*-value)
Original *suicide* values	0.02345	1.875×10^{-5}
Logarithmic *suicide* values	0.28360	0.7268

The later issue is to examine the effect of such a transformation on the dependence from the *income* variable, both from a descriptive and an inferential point of view by comparing results from OLS and QR.

First, the logarithmic transformation allows the *suicide* heteroskedasticity to be reduced and normal distributed residuals to be obtained, as shown in Table 4.2, using White's general test for heteroskedasticity (White 1980) and the Shapiro–Wilk normality test (Shapiro and Wilk 1965, 1968). The two tests are applied to two OLS regressions using *income* as an explanatory variable and *suicide* or its logarithmic transformation as a dependent variable. Moving from the original *suicide* variable to its logarithmic transformation, the normality and homoskedasticity assumptions cannot be refused.

Visual inspection of the residuals using a Q-Q plot confirms the effect of the logarithmic transformation on OLS regression (Figure 4.3).

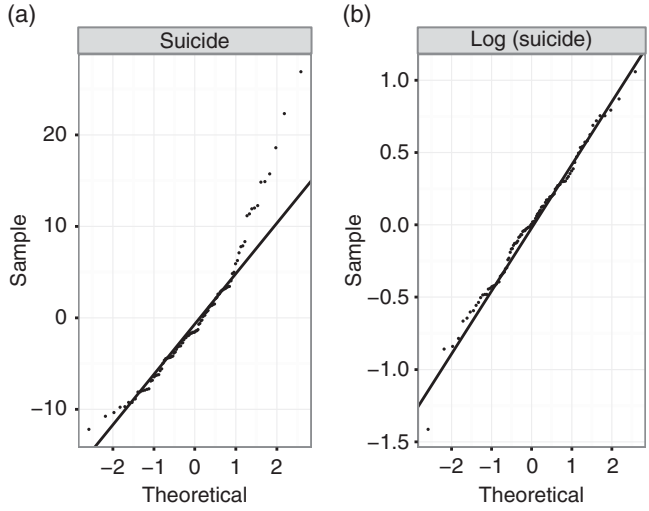

Figure 4.3 Residual Q-Q plot without (a) and with (b) the logarithmic transformation of the suicide*-dependent variable. Moving from the original* suicide *to its logarithmic transformation, the normality assumption becomes more evident.*

Table 4.3 OLS and QR estimates (θ=0.1, 0.5, 0.9) with and without a logarithmic transformation of the dependent variable. The logarithm of the predicted value obtained through a QR on the original dependent variable is equal to the predicted value obtained estimating the same model but using the logarithmic transformation of the dependent variable, whatever the value of the explanatory variable. OLS regression does not benefit this property.

Model	$E(\mathbf{y}\|\mathbf{x})$	$E(log(\mathbf{y})\|\mathbf{x})$	$Q_{0.5}(\mathbf{y}\|\mathbf{x})$	$Q_{0.5}(log(\mathbf{y})\|\mathbf{x})$
Constant	3.111	1.817	1.510	1.881
Coefficient	0.382	0.025	0.383	0.023
Model	$Q_{0.1}(\mathbf{y}\|\mathbf{x})$	$Q_{0.1}(log(\mathbf{y})\|\mathbf{x})$	$Q_{0.9}(\mathbf{y}\|\mathbf{x})$	$Q_{0.9}(log(\mathbf{y})\|\mathbf{x})$
Constant	1.319	1.358	−1.256	0.771
Coefficient	0.225	0.024	2.224	0.029

If the logarithmic transformation and, consequently, the normality of the dependent variable profits both the OLS and the QR, the consequences for the values of the parameter estimates are different in the two methods.

In Table 4.3, OLS and some QR (θ=0.1, 0.5, 0.9) estimates with and without a logarithmic transformation of the dependent variable are shown. The equivariance to monotone transformations can easily be numerically proved by fixing a value for the *income* variable. For example, if the *income* is equal to 28 (million old Italian lire), the predicted value of the model with $\theta = 0.1$ and without a logarithmic

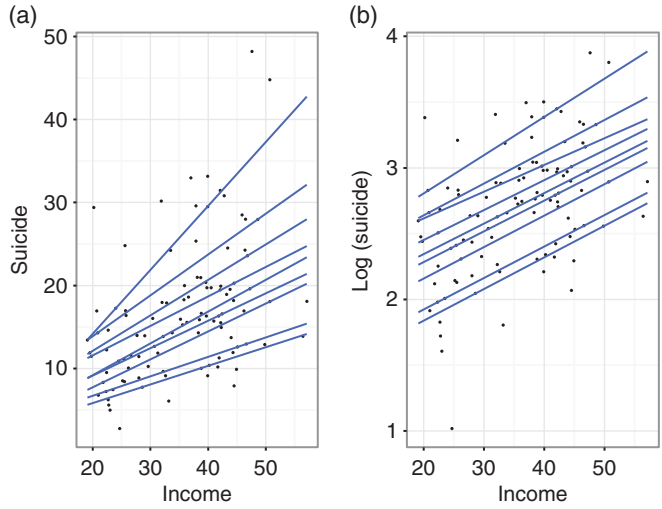

Figure 4.4 Scatter plot and QR lines without (a) and with (b) a logarithmic transformation of suicide *(quantiles* θ *= [0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9]). The scale effect shown by the original* suicide *variable disappears in the case of its logarithmic transformation where a pure location effect remains.*

transformation of the dependent variable is 7.6. Its log value (2.0) corresponds to the predicted value of the same model with a logarithmic transformation of the dependent variable:

$$\begin{cases} Q_{0.1}(\widehat{suicide}|income = 28) = 1.319 + 28 \times 0.225 = 7.61 \\ Q_{0.1}(\widehat{log(suicide)}|income = 28) = 1.358 + 28 \times 0.024 = 2.03. \end{cases}$$

Obviously, taking the exponential of 2.03 leads to the predicted value of the model $Q_{0.1}(\mathbf{y}|\mathbf{x})$.

The interpretation of the coefficients in a logarithmic transformation can also be done in terms of the percentage change of a given explanatory variable: $100 \times (e^{log(\beta(\theta))} - 1)$ in the θ-th quantile of the response variable. For example, a unitary increase of *income* causes a relative effect on the *rate of suicide* equal to 2.4% in the model where $\theta = 0.1$.

The effect of the logarithmic transformation on the results of QR is evident in Figure 4.4. Moving from Figure 4.4(a) to (b), the QR lines become almost parallel, a sign of a pure location effect showing that *income* exerts a constant percentage change in all the *suicide* values (Koenker and Xiao 2002).

To describe the effect played by a logarithmic transformation on the inference results, a random sample of 60 units is extracted from the *IlSole24Ore* dataset.

Let us consider a very simple inferential problem represented by the construction of a confidence interval for the population mean. If the parameter of interest is the

Table 4.4 Ninety-five percent confidence intervals for the *suicide* mean with and without a logarithmic transformation of the dependent variable. Transforming back the extremes of the confidence interval obtained using the logarithm of the dependent variable can provide an interval unable to include the population mean.

	Sample mean	95% Confidence interval	
Original *suicide* values	16.042	13.851	18.232
Logarithmic *suicide* values	2.655	2.526	2.784

suicide mean, 95% confidence intervals can be obtained with and without the logarithmic transformation (Table 4.4). In this simple example, the population mean is known (16.652) and allows us to prove that, when transformed back to the original scale, the confidence interval obtained with the logarithmic *suicide* values produces results contrary to those at the original scale. Basically, when the population mean falls inside the interval constructed for the original values (13.851; 18.232), it falls outside the interval derived from transforming back the extremes constructed for the logarithmic values, where the exponential of 2.526 equals 12.503 and the exponential of 2.784 equals 16.184.

The logarithmic transformation might be very hazardous in terms of the inference results of an OLS regression (Manning 1998) whereas it may aid the statistical inference of QR (Cade and Noon 2003).

To evaluate the consequences of a logarithmic transformation for the parameter inference in more detail, both in OLS and QR analysis, the dummy *temperature* explanatory variable is again considered. Figure 4.5 and Figure 4.6 show the *suicide* distribution in the two groups of provinces with and without the logarithmic transformation.

In OLS regression, inference on a transformed dependent variable should be interpreted very cautiously because the evaluation of the significance of the parameter values can lead to different conclusions with and without the use of the transformation. For example, looking at the *p*-values of the parameters in Table 4.5, it is evident they are quite different, whereas those derived from QR (Table 4.6) are almost the same. Consequently, inference on the QR results is not affected by a monotone transformation, and it can even be improved (Chen 2005).

To describe the QR equivariance to a different monotone transformation capable of dealing with negative skewness (Manning 1998), the *obese* dataset is considered (Santamaria *et al.* 2011). It refers to a sample of 171 obese individuals. The aim of this dataset is to analyze if and how the *forced vital capacity* (FVC) depends on the *body mass index* (BMI).

The distribution of the FVC-dependent variable is clearly asymmetric [Figure 4.7(a)], and it can be a possible cause of the heteroskedasticity between the FVC and the BMI [Figure 4.7(b)]. In these kinds of situations, before performing a dependence analysis, it is advisable to properly transform the dependent variable to reduce its asymmetry, or even to bring it closer to a normal distributed variable.

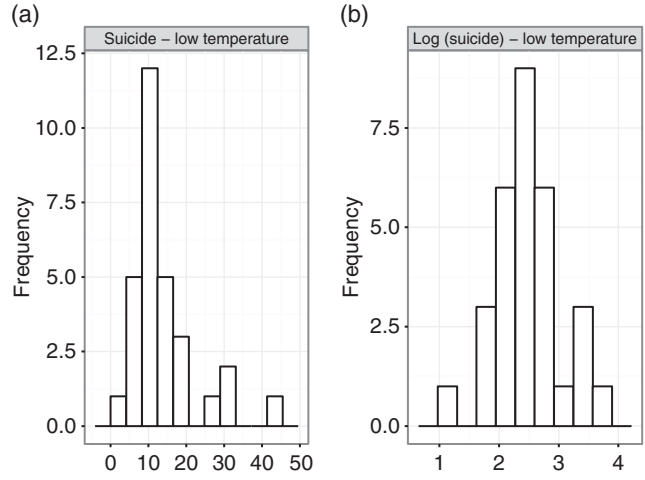

Figure 4.5 Histogram of the suicide *for provinces with a low* temperature *with (b) and without (a) the logarithmic transformation. Moving from the original* suicide *variable to its logarithmic transformation, the normality assumption becomes closer.*

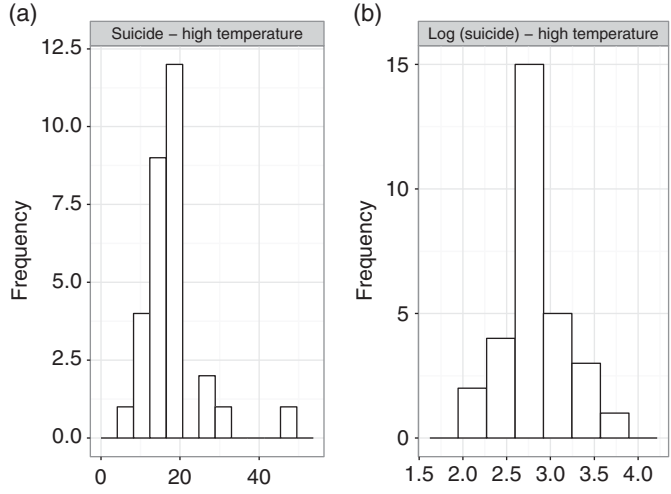

Figure 4.6 Histogram of the suicide *for the provinces with a high* temperature *with (b) and without (a) the logarithmic transformation. Moving from the original* suicide *variable to its logarithmic transformation, the normality assumption becomes closer.*

The Box–Cox method (Box and Cox 1964) is typically used to suggest the best transformation in cases of skewness. This method is based on a family of power transformations of the $\mathbf{x}$ variable, which is obtained by varying the λ parameter:

$$\mathbf{x}_{trans} = \frac{\mathbf{x}^{\lambda} - 1}{\lambda}. \tag{4.13}$$

Table 4.5 OLS parameter values and inference results with and without a logarithmic transformation of the dependent variable. The obtained p-values are rather different.

		Estimate	Standard error	p-value
Original *suicide* values	Intercept	13.934	1.512	0.000
	Temperature=high	4.215	2.138	0.053
Logarithmic *suicide* values	Intercept	2.475	0.086	0.000
	Temperature=high	0.359	0.121	0.004

Table 4.6 QR parameter values and inference results ($\theta = 0.5$) with and without a logarithmic transformation of the dependent variable. The obtained p-values are rather similar.

		Estimate	Standard error	p-value
Original *suicide* values	Intercept	11.600	1.082	0.0000
	Temperature=high	5.320	1.437	0.0004
Logarithmic *suicide* values	Intercept	2.451	0.083	0.0000
	Temperature=high	0.377	0.097	0.0002

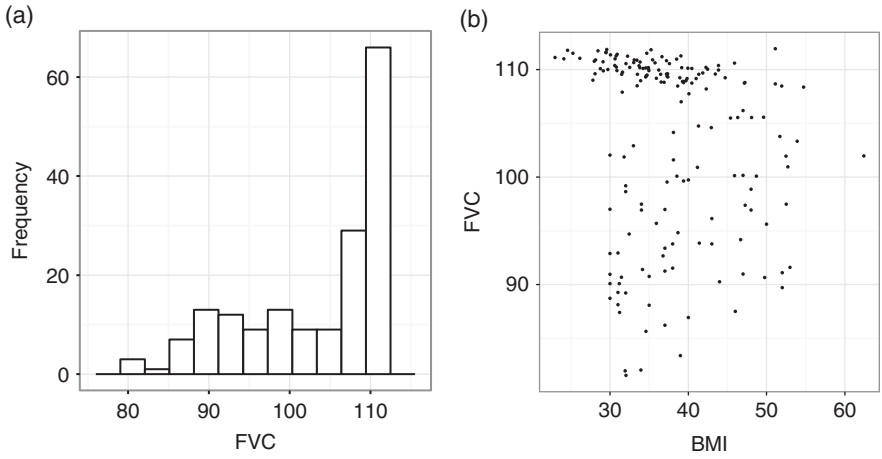

Figure 4.7 FVC distribution (a) and scatter plot of FVC and BMI (b). The dependent variable shows a negative skewness. As the BMI increases, the variability of the FVC also increases, a sign of a heteroskedastic relationship between the two variables.

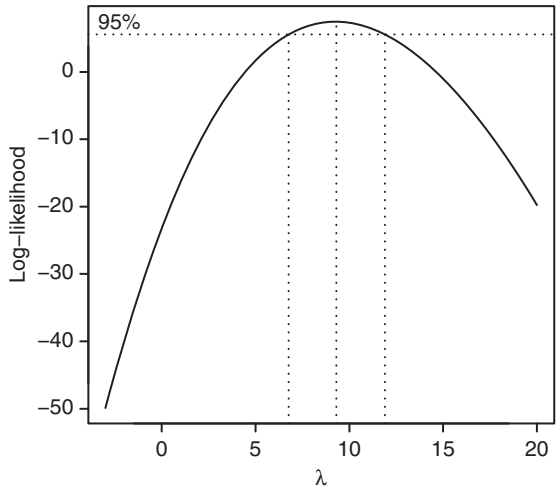

Figure 4.8 Box–Cox log-likelihood function. The log-likelihood function is represented on the vertical axis for different values of the λ parameter.

Usually λ varies in the range [−3; +3], and the best value is the one able to maximize the so-called log-likelihood function:

$$L = \frac{\nu}{2} ln(s^2_{trans}) + (\lambda - 1)\frac{\nu}{n}\sum_{i=1}^{n} ln\, x_i. \tag{4.14}$$

where ν equals $n - 1$ because it represents the number of degrees of freedom, and s^2_{trans} is the variance of the transformed variable.

The Box–Cox method applied to the FVC variable suggests that the FVC should be increased to a power equal to 9.3 (Figure 4.8) because it guarantees the maximum value of the log-likelihood function. Such transformation allows the asymmetry and heteroskedasticity to be reduced, but it is not able to transform FVC into a normally distributed variable. Consequently, applying an OLS regression on the powered FVC poses problems, both from a descriptive and an inferential point of view.

In the case of an increase in the dependent variable to a given power, the QR equivariance property can be numerically proven, as in the case of the logarithmic transformation (Table 4.3).

In Table 4.7, the OLS and the median QR estimates with and without a power transformation of the dependent variable are shown. Fixing a value for the BMI variable, for example 46, the predicted value of the quantile model with $\theta = 0.5$, and the original values of the dependent variable is 1.06.[1] Its power to 9.3 (1.66) corresponds to the predicted value in the same model, with an increase of the dependent variable

[1] To simplify the results interpretation, the original values of *FVC* have been divided by 100.

Table 4.7 OLS and QR estimates with and without an increase in the FVC to the power of 9.3. QR allows to easily move from results obtained using the transformed dependent variable to those obtained on the original data.

Model	$E(\mathbf{y}\|\mathbf{x})$	$E(\mathbf{y}^{9.3}\|\mathbf{x})$	$Q_{0.5}(\mathbf{y}\|\mathbf{x})$	$Q_{0.5}(\mathbf{y}^{9.3}\|\mathbf{x})$
Constant	1.094	2.771	1.187	3.867
Coefficient	−0.001	−0.029	−0.003	−0.046

to the power of 9.3. Obviously, taking the 9.3th root of 1.66 leads to the predicted value of the model $Q_{0.5}(\mathbf{y}|\mathbf{x})$.

4.2 Response conditional density estimations

Density estimation is a useful statistical tool in exploring and analyzing data (Silverman 1986; Hyndman *et al.* 1996). In QR, the construction of an estimate of the density function of the response variable can be realized using two different approaches: simulating different conditional densities of the response variable by conditioning on specific values of the explanatory variables and estimating the whole conditional response distribution by exploiting the entire quantile process.

4.2.1 The case of different scenario simulations

The response variable conditional density for different scenarios of the explanatory variables can be obtained by starting from the set of estimated QR coefficients and conditioning the explanatory variables to values of interest. In this section, the explanatory and informative capabilities of the conditional densities are described in three different contexts: where there is one explanatory variable, where there is more than one explanatory variable and where the values to condition the explanatory variables are artificial. The first case is described using the *IlSole24Ore* dataset and it aims to highlight the principal differences between the main approaches to density estimation in QR: histograms and the kernel method. The *degree* dataset is more complex because it is possible to condition to the values of many explanatory variables. Using this dataset, it is possible to define typologies and to explore differences among groups of units in terms of the conditional density of the response variable. Finally, using the *satisfaction* dataset, a different study is proposed of the effect played by fixed combinations of the explanatory variables on the response variable conditional density. In particular, the aim is to hypothesize the consequences for the response variable in cases of extreme or particular typologies of units.

First, the *IlSole24Ore* dataset is used to explore the response variable conditional density. The estimated empirical quantile function for *suicide* can be analyzed by dividing the Italian provinces according to the *income* values in *poor* provinces

where values of *income* equal to 22.46 correspond to the 10th percentile of the variable distribution and in *rich* provinces where values of *income* equal to 46.48 correspond to the 90th percentile. It follows that for any quantile of interest θ, it is possible to obtain two conditional distributions of the response variable

$$\begin{cases} Q_\theta^{poor}\left(\widehat{suicide}|income = 22.46\right) \\ Q_\theta^{rich}\left(\widehat{suicide}|income = 46.48\right) \end{cases}$$

To reconstruct the whole conditioned distribution of the response variable, the QR is estimated for the solutions of all θ values.

Once the estimated response values are obtained, the next step is to adopt the best statistical tools to explore features and differences between the two distributions. To this end, one of the most widespread approaches is the analysis of the variable probability density function because it allows many properties of the variable of interest to be highlighted. Standard density estimation methods include parametric and nonparametric approaches. In the following, the second approach is privileged because it does not require rigid assumptions about the distribution of the observed data, and the data determine the shape of the density. Histograms are the oldest and most widely used density estimators (Silverman 1986), and the kernel method is well appreciated for its statistical properties and for providing a smoothed representation.

By comparing the histogram of the *suicide* density for the whole set of units with the units split into the *poor* and *rich* provinces (Figure 4.9), it can be seen that the modes of the *suicide* distribution around the *income* values equal to 10 and 20 can

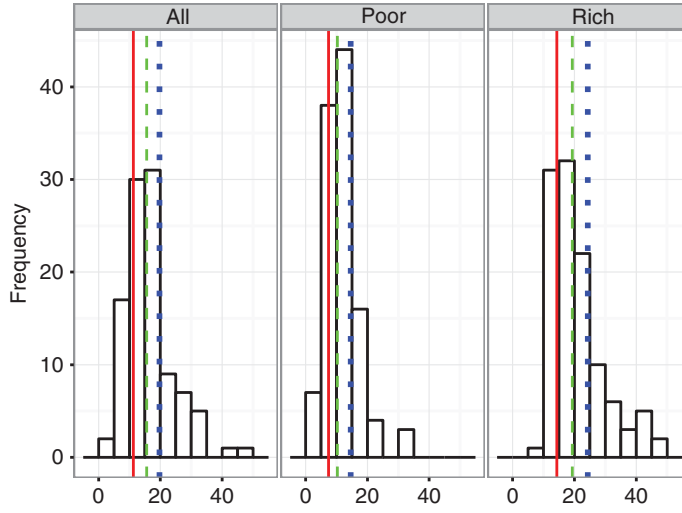

Figure 4.9 Histograms of the suicide *density in the whole dataset (left) and the* poor *(center) and the* rich *(right) provinces. The distribution in the whole dataset is multimodal, and each mode characterizes a group of provinces according to income.*

be explained by the presence of the two groups of provinces (*poor* and *rich*) with different features in the *suicide* density features. Both histograms of the *poor* and *rich* provinces show an asymmetric bimodal shape, but the second is shifted towards higher *suicide* values. Consequently, the main statistics shown on the vertical axis (solid line for the first quartile, dashed line for the median and dotted line for the third quartile) are also shifted.

Histograms provide different pictures of data by varying the starting point of the first interval and the number of intervals. The first parameter influences the form of the histogram, and the second one affects the degree of smoothing.

Several methods can be adopted to define the number of intervals, but they are beyond the scope of this section, which aims to describe how much informative content is held in a response conditional density resulted from a QR. It is important to highlight differences derived from a low or a high number of intervals, even if such numbers are arbitrary. Figure 4.10 and Figure 4.11 show the histogram with a bin width equal to 1 beside the histogram depicted in Figure 4.9 (with a bin width equal to 5). Decreasing the bin width reveals a less asymmetric density for *poor* provinces and a mode less than 10, different from the one in the histogram on the left. Regarding *rich* provinces, the histogram on the right seems under-smoothed (too bumpy), but it shows several modes.

To obtain more smoothed densities, an alternative density estimation procedure represented by the kernel can be used:

$$\hat{f}(\mathbf{y}) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\mathbf{y} - y_i}{n}\right), \tag{4.15}$$

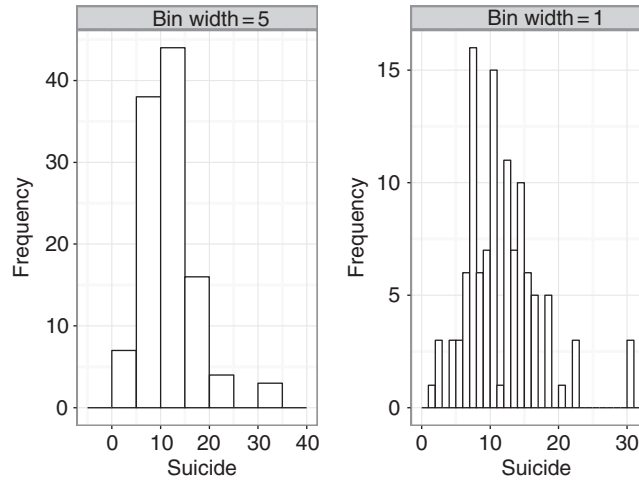

Figure 4.10 Histograms of the suicide *distribution conditioned to the value of the* income *in the 10th percentile. For* poor *provinces, a less asymmetric distribution emerges, increasing the number of breaks.*

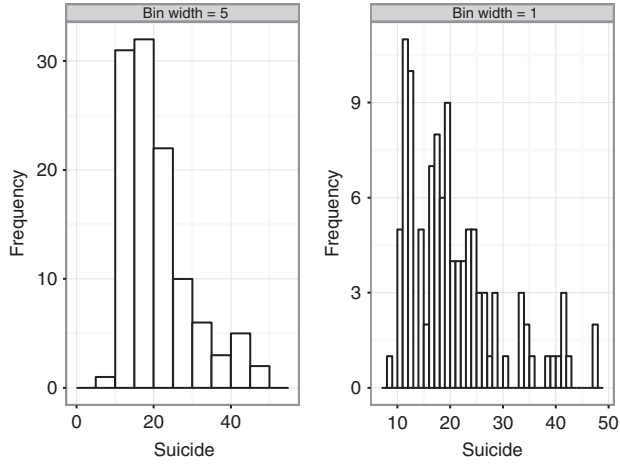

Figure 4.11 Histograms of the suicide *distribution conditioned to the value of the* income *in the 90th percentile. For* rich *provinces, as the number of breaks increases, the distribution becomes even more multimodal.*

where $y_1, \ldots, y_n$ represents the sample of n observations, $\hat{f}(\mathbf{y})$ the density estimator, h the width of each interval and $K(.)$ the kernel function (Ahamada and Flachaire 2011).

It is well known that the kernel function has to satisfy all the properties of a density function and that Gaussian and Epanechnikov kernels are the most widely used. The literature suggests that the effect played by different kernel functions is not as relevant as consequences related to different choices of the h parameter. This parameter is known as the *smoothing* or *bandwidth* parameter because it influences the degree of smoothness. For small values of h, spurious structures may be visible but at the cost of higher variance estimators. In contrast, large values of h can cause an over-smoothing of the density function. A classical approach to choose the value of the bandwidth parameter is the graphical inspection of different alternative choices.

In Figure 4.12 and Figure 4.13, a Gaussian distribution has been used as the kernel function, even if this may not be the true distribution. Comparing the four graphs obtained by varying the bandwidth parameter shows that the density constructed with a bandwidth of around 1.8 is the best in terms of visualized information.

Such a value is suggested by the following widespread rule of thumb proposed by Silverman (1986) to identify the optimal value for the smoothing parameter:

$$h_{opt} = 0.9min\left(\hat{\sigma}; \frac{\hat{q}_3 - \hat{q}_1}{1.349}\right) n^{-\frac{1}{5}}, \tag{4.16}$$

where $\hat{\sigma}$, $\hat{q}_1$ and $\hat{q}_3$ denote, respectively, the standard deviation, the first quartile and the third quartile of the data.

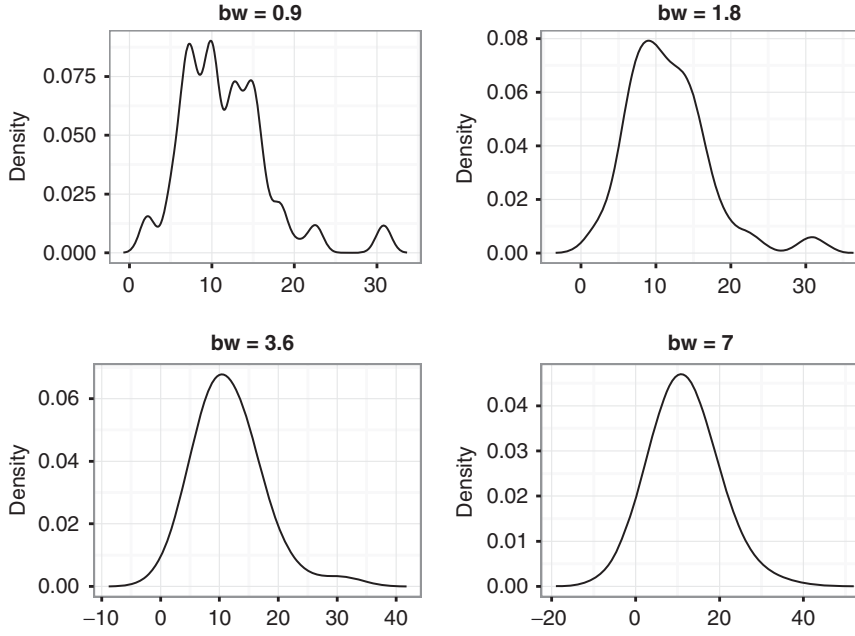

Figure 4.12 Kernel density estimations of the suicide *distribution conditioned to the value of the* income *in the 10th percentile and for different values of the bandwidth (bw) parameter. A good compromise between the extreme analyticity deriving from low bandwidth values and the over-smoothing in cases of high bandwidth values is obtained when the bandwidth equals 1.8.*

The classical kernel density estimation is not the best solution in a QR problem where the considered distribution is typically asymmetric and multimodal (Van Kerm 2003). In such cases, a constant bandwidth can cause over-smoothing of the mode part of the distribution and under-smoothing in the tails, thus providing estimates with reduced variance in the tails and reduced bias in the modal part. Instead, adaptive methods, such as adaptive kernel estimation, can be used that enable the bandwidth to be varied on the basis of the distribution concentration:

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\lambda_i}K\left(\frac{\mathbf{y}-y_i}{h\lambda_i}\right). \tag{4.17}$$

The adaptive kernel estimator in Equation (4.17) includes a further parameter, λ_i, which is considered a local bandwidth factor because it varies according to the concentration of the distribution in each point (Abramson 1982):

$$\lambda_i = \left[\frac{g}{\tilde{f}(y_i)}\right]^{\alpha} \tag{4.18}$$

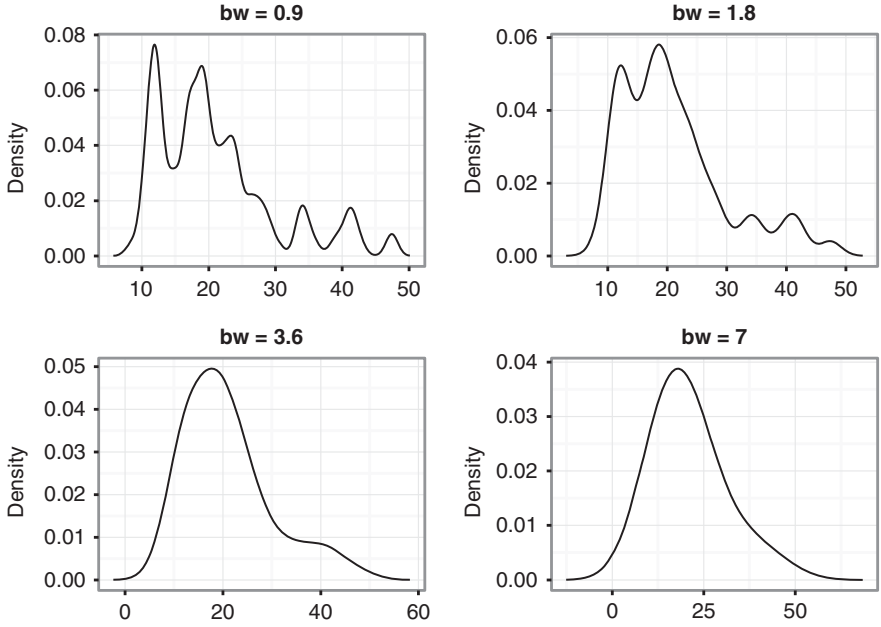

Figure 4.13 Kernel density estimations of the suicide *distribution conditioned to the value of the* income *in the 90th percentile and for different values of the bandwidth parameter. A good compromise between the extreme analyticity deriving from low bandwidth values and the over-smoothing in cases of high bandwidth values is that obtained when the bandwidth equals 1.8.*

where $\tilde{f}(y_i)$ is a pilot function, namely that used for a first estimation of the density, g is its derivative and α is a sensitivity parameter, which varies from 0 to 1.

Figure 4.14 and Figure 4.15 compare the classical kernel density with the adaptive density both for *poor* and *rich* provinces, setting the bandwidth h equal to the rule of thumb proposed by Silverman (4.16), the kernel function as normal and the sensitivity parameter α equal to 0.5. Such a value results in reduced bias in the estimation of the density function (Abramson 1982).

Adaptive kernel estimation improves the estimates of the *poor* and the *rich* conditioned distributions, both in the main part of the distribution and in the tail where the bumps are reduced.

The analysis of the conditional density response variable can be expanded by fixing interest values for more than one explanatory variable. With this aim in mind, let us consider the *degree* dataset used to evaluate how and if the student features affect the outcome of university careers (*degree mark*) (Davino and Vistocco 2008a). The analysis is performed on the data gathered through a survey conducted on a random sample of 685 students who had graduated from the University of Macerata, which is located in the Italian Marche region. The survey was conducted in 2007, and it included students who graduated 2–5 years earlier. The following features of

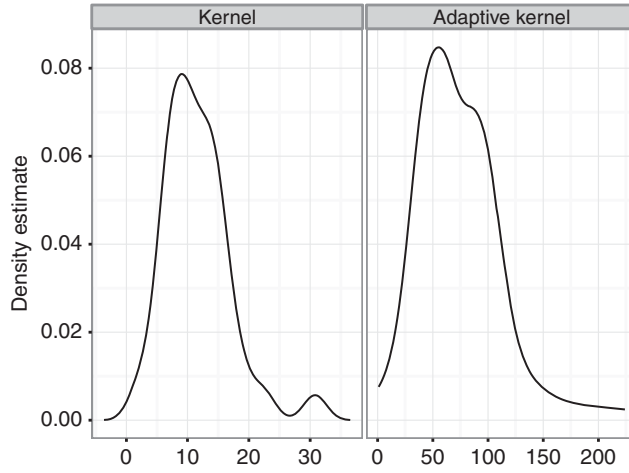

Figure 4.14 Kernel and adaptive density estimations of the suicide *distribution conditioned to the value of the* income *in the first quartile. Using the bandwidth parameter suggested by the rule of thumb of Silverman and a sensitivity parameter equal to 0.5, the adaptive kernel density estimation offers more informative content than the kernel density estimation.*

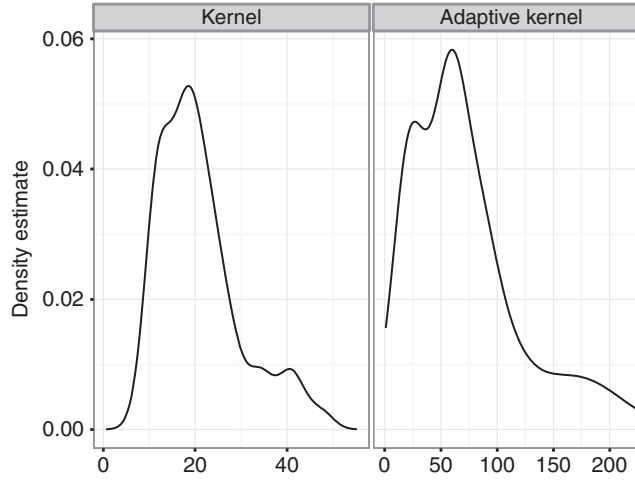

Figure 4.15 Kernel and adaptive density estimations of the suicide *distribution conditioned to the value of the* income *in the third quartile. Using the bandwidth parameter suggested by the rule of thumb of Silverman and a sensitivity parameter equal to 0.5, the adaptive kernel density estimation offers more informative content than the kernel density estimation.*

the student profile were observed and considered as explanatory variables: *gender*, *place of residence* during university education (Macerata and its province, Marche region, outside Marche), *course attendance* (no attendance, regular), *foreign experience* (yes, no), *working condition* (full-time student, working student), *number of years to obtain a degree*, and *diploma mark*.

The *degree mark* conditional density can be explored according to the *number of years to obtain a degree* (4, 5, 6, 7, 8, 9). The six groups of students share the following features for the other explanatory variables: female, place of residence in Macerata, regular course attendance, foreign experience, working student, and lowest diploma mark. Figure 4.16 shows the six conditional distributions of the *degree mark* along with the corresponding quartiles (vertical segments). The conditional distributions differ according to the *number of years to obtain a degree* above all in the values of the first quartile.

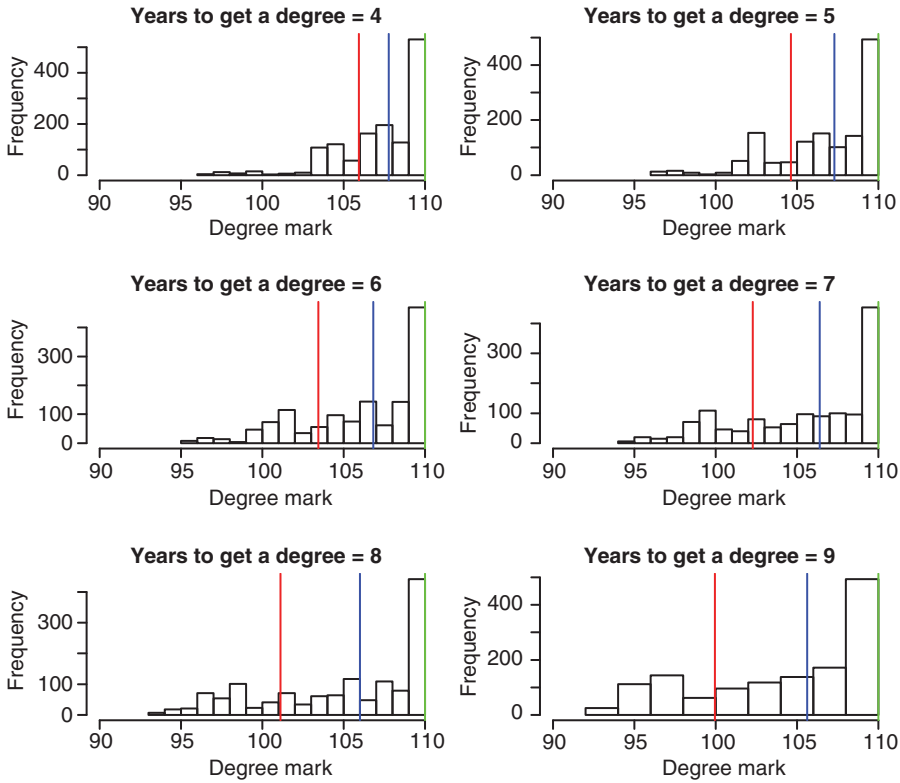

Figure 4.16 Histograms of the conditional distribution of the degree mark (sex=female, place of residence=Macerata, course attendance=regular, diploma mark=lowest) for the number of years to obtain a degree. Vertical segments represent quartiles. Moving from regular to slower students, the conditional density of the degree mark changes.

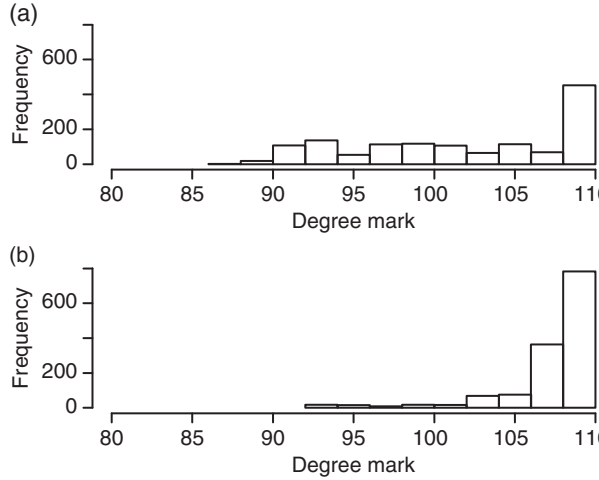

Figure 4.17 Histograms of the conditional distribution of the degree mark for (a) working students (sex=male, place of residence=outside Marche, course attendance=no, diploma mark=lowest) and (b) regular students (sex=female, place of residence=Macerata, course attendance=regular, diploma mark=highest). The degree mark tends to show higher variability and extreme low values in the group of working students.

More pronounced differences appear distinguishing working from regular students and fixing the other explanatory variables to males, place of residence outside Marche, no course attendance and lowest diploma mark for working students and to males, place of residence in Macerata, regular course attendance and highest diploma mark for full-time students. The group of working students [Figure 4.17(a)] shows a higher variability and extreme low values in the conditioned *degree mark* distribution.

A further example of the usefulness of the conditional density estimates is provided using extreme profiles of units obtained assigning hypothetical values to the explanatory variables. The *satisfaction* dataset lends itself to this type of analysis. The evaluation is based on the analysis of data derived from a survey that is undertaken each year at the University of Macerata (Italy) to monitor student satisfaction with respect to the courses they attend. The course features considered are: *course design* (exam conditions, course scheduling, availability of the teacher, co-ordination with other courses), *teaching and study* (preliminary knowledge required, the ability of the teacher to attract interest, clarity of the teacher, study load, suitable materials for studying, explanation of topics with respect to time, interaction with the teacher during the lessons) and *infrastructures* (comfort of classrooms). The dependent variable is the *overall student satisfaction*.

The aim of the analysis is to evaluate how much judgments about the different features of a course affect the *overall satisfaction* (Davino and Vistocco 2008b).

It is worth specifying that all the variables have been observed on a four-level ordinal scale (definitely unsatisfied, unsatisfied, satisfied, definitely satisfied) but

before the analysis phase a direct quantification is used by assigning to the categories the following scores: 2, 5, 7, 10. There is consensus among the Italian scientific community on such a quantification approach, and it is accepted by the National Committee for University System Evaluation (CNVSU) (Chiandotto and Gola 1999) in the specific context of the evaluation of student satisfaction.

To explore the level of concordance of the overall satisfaction with the evaluation of all the features of a course, extreme profiles of units are considered. These extreme and hypothetical profiles are defined using uniform judgments for all the considered features. For example unsatisfied profiles correspond to judgments equal to 2 for all the variables, and they represent the explanatory variable values conditioning the overall satisfaction. The conditioned density distributions of the dependent variable (Figure 4.18) reveal that the distribution is more concentrated in cases of high scores than in cases of low scores, showing a strong correlation between the

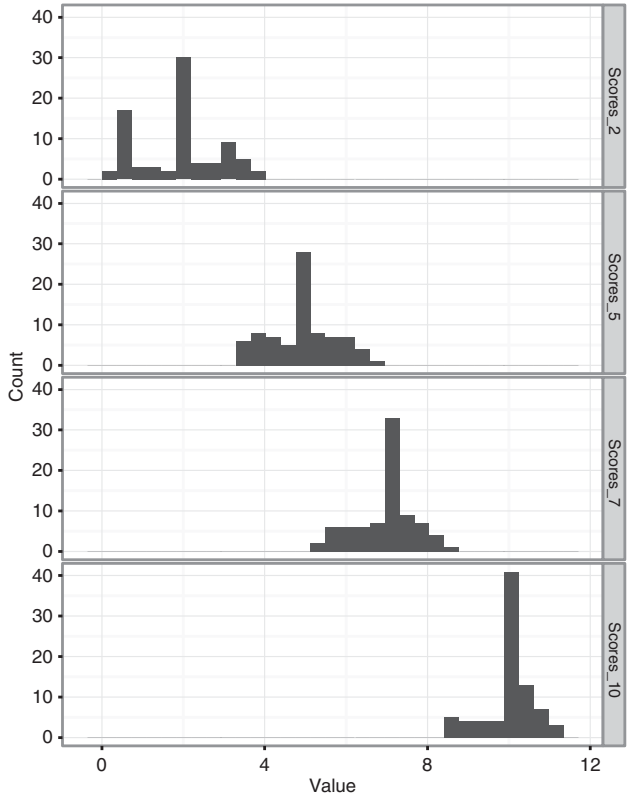

Figure 4.18 Histogram of student satisfaction conditional densities corresponding to extreme uniform profiles, namely hypothetical students making uniform judgments of all the explanatory variables. The distribution is more concentrated than in cases of low scores, showing a strong correlation between the final score and the feature scores in cases of satisfied students.

final score and the feature scores in cases of satisfied students. On the contrary, the final score of unsatisfied students is more heterogeneous. It is worth noting that overlapping regions are present both between profile 1 (Score_2) and profile 2 (Score_5) and between profile 2 (Score_5) and profile 3 (Score_7). Such behavior suggests that suitable policies could allow profile 1 students to move to profile 2 and profile 2 students to move to profile 3.

4.2.2 The case of the response variable reconstruction

The approach presented in Section 4.2.1 is also suitable for estimating the whole conditional response distribution derived from exploiting the QR process.

While predicted values are uniquely calculated in an OLS regression, they vary in QR according to the desired quantile.

To provide a unique prediction vector, it is possible to identify, for each unit, the QR model best able to estimate the response variable. Let us consider the QR model for a given conditional quantile θ:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}(\theta). \tag{4.19}$$

The generic element of the $\hat{\mathbf{Y}}(n \times k)$ matrix is the estimate of the response variable in correspondence to the i-th unit according to the θ-th quantile. The best estimate for each unit is the one that minimizes the difference between the observed and the estimated value, for each of the k models (Davino and Vistocco 2008a):

$$\theta_i : argmin_{\theta=1,\ldots,k} \left| y_i - \hat{y}_i(\theta) \right|. \tag{4.20}$$

Once θ_i is identified for each unit i, the dependent variable is reconstructed assigning to each unit the $\hat{y}$ estimated at the corresponding quantile.

Referring again to the *IlSole24Ore* dataset, Figure 4.19 displays the smooth density estimate of the original *suicide* variable and the estimates of the corresponding predictions obtained through an OLS and QR. The capability of QR to reconstruct almost perfectly the distribution of the response variable is clear.

4.3 Validation of the model

4.3.1 Goodness of fit

The assessment of goodness of fit for the QR model exploits the general idea leading to the typical R^2 goodness of fit index in classical regression analysis (Koenker and Machado 1999).

Among the different available formulations of the R^2 index, the formulation expressed in terms of the complement to 1 of the ratio between the residual sum of squares and the total sum of squares of the dependent variable (Gujarati 2003) allows us to derive the corresponding index for QR. The latter can be obtained by

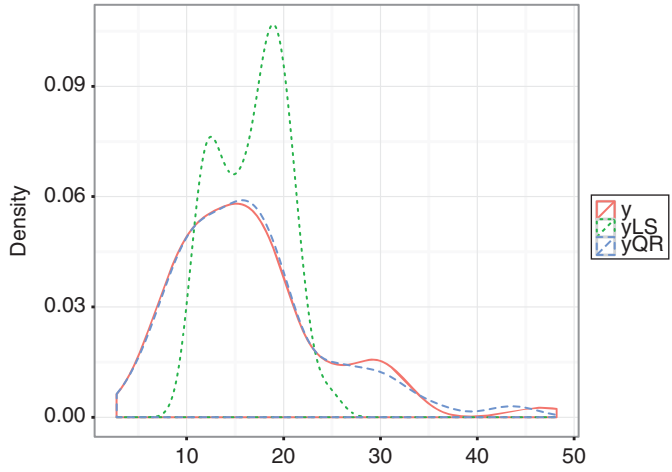

Figure 4.19 Density estimation of the original response variable (y) and estimation of the predictions obtained through OLS (yLS) and QR (yQR). The distribution of the response variable can be reconstructed almost perfectly with QR.

taking into account that QR differs from OLS regression because it is based on the minimization of an absolute weighted sum (not an unweighted sum of squares as in OLS). In addition, this approach aims to provide an estimation of the quantile conditional distribution of the dependent variable (not an expected conditional distribution as in OLS). It follows that the equivalent of the residual sum of squares is, for each considered quantile θ, the residual absolute sum of weighted differences between the observed dependent variable and the estimated quantile conditional distribution.

For the simplest regression model with one explanatory variable:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}. \tag{4.21}$$

The residual absolute sum of weighted differences is the corresponding minimizer:

$$\begin{aligned} RASW_\theta = &\sum_{y_i \geq \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} \theta \left| y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i \right| + \\ &\sum_{y_i < \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} (1-\theta) \left| y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i \right|. \end{aligned} \tag{4.22}$$

The equivalent of the total sum of squares of the dependent variable is, for each considered quantile θ, the total absolute sum of weighted differences between the observed dependent variable and the estimated quantile:

$$TASW_\theta = \sum_{y_i \geq \theta} \theta \left| y_i - \hat{\theta} \right| + \sum_{y_i < \theta} (1-\theta) \left| y_i - \hat{\theta} \right|. \tag{4.23}$$

The obtained $pseudoR^2$ can be considered as an index comparing the residual absolute sum of weighted differences using the selected model with the residual absolute sum of weighted differences using a model with only the intercept. The obtained $pseudoR^2$ can be computed as follows:

$$pseudoR^2_\theta = 1 - \frac{RASW_\theta}{TASW_\theta}. \tag{4.24}$$

As $RASW_\theta$ is always less than $TASW_\theta$, the $pseudoR^2_\theta$ ranges between 0 and 1. It is worth noting that the index cannot be considered a measure of the goodness of fit of the whole model because it is related to a given quantile. In practice, for each considered quantile, the corresponding $pseudoR^2$ can be evaluated at a local level, thereby indicating whether the presence of the covariates influences the considered quantile.

In Table 4.8, the previously defined differences and the corresponding $pseudoR^2$ are listed for a given number of models estimated using the *IlSole24ore* dataset with *suicide* as a dependent variable and *income* as an explanatory variable.

The $pseudoR^2$ can also be used to assess the model with the best goodness of fit between nested models. If two nested models are considered, the comparison between the corresponding $pseudoR^2$ can be used to evaluate whether it is advantageous to move from a restricted to a more complex model. Any eventual improvement can be quantified through the so-called $relativeR^2$. This is derived from a different interpretation of the $pseudoR^2$, where *RASW* and *TASW* represent the residual absolute sum of weighted differences between the observed dependent variable and the estimated quantile conditional distribution, respectively, in the more complex model and in the simplest one (Hao and Naiman 2007).

Table 4.8 Weighted differences and $pseudoR^2$ for a given number of models. For each selected quantile, a $pseudoR^2$ can be associated with the relative model.

Quantiles	$RASW_\theta$	$TASW_\theta$	$pseudoR^2_\theta$
0.1	183.84	212.76	0.14
0.2	324.97	365.70	0.11
0.3	432.92	482.03	0.10
0.4	497.41	560.90	0.11
0.5	536.71	603.78	0.11
0.6	548.06	612.58	0.10
0.7	523.34	588.65	0.11
0.8	458.45	523.67	0.12
0.9	317.07	354.30	0.10

4.3.2 Resampling methods

Resampling methods can represent a valid alternative to the asymptotic inference described in Chapter 3 because they allow estimation of parameter standard errors without requiring any assumption in relation to the error distribution (Gould 1992).

The drawback of using resampling methods (Efron and Tibshirani 1998) is their computational cost, even if the cost is constantly being reduced by the development of modern technologies. Moreover, although bootstrap estimates are unbiased, they present different sources of variability because they are based on one single sample from a given population (sample variability) and on a finite number of replications (bootstrap resampling variability).

Several contributions in the literature suggest bootstrap as the most suitable resampling method in QR analysis. A number of these are discussed and compared in this subsection:

- *xy*-pair method or design matrix bootstrap (Kocherginsky 2003);
- method based on pivotal estimating functions (Parzen *et al.* 1994);
- Markov chain marginal bootstrap (He and Hu 2002; Kocherginsky 2003; Kocherginsky *et al.* 2005).

To appreciate the differences and peculiarities of bootstrap methods, homogeneous and heterogeneous variance regression models are considered. In particular, two of the simulated datasets described in Chapter 2, Section 2.2, are generated with the following characteristics: a sample of $n = 10\,000$ observations is extracted from $model_1 \rightarrow \mathbf{y}^{(1)} = 1 + 2\mathbf{x} + \mathbf{e}$ (homogeneous error model) and from $model_6 \rightarrow \mathbf{y}^{(6)} = 1 + 2\mathbf{x} + (1 + \mathbf{x})\mathbf{e}$ (heterogeneous error model), where $\mathbf{x} \sim N(10; 1)$ and $\mathbf{e} \sim N(0; 1)$.

4.3.2.1 *xy*-pair method or design matrix bootstrap

Let us consider the simplest QR model with one explanatory variable and n observations:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1(\theta)\mathbf{x}. \tag{4.25}$$

The *xy*-pair method or design matrix bootstrap consists of constructing a given number of samples (B), usually with the same size as the original dataset, where each sample is obtained by a random sampling procedure with replacement from the original dataset. The resampling procedure is simultaneously applied to the $\mathbf{x}$ and $\mathbf{y}$ vectors. B QRs are performed on the bootstrap samples, and a vector of the parameter estimates is retained for each quantile of interest (Figure 4.20).

Whatever statistic of interest can be calculated on the previous vector of parameter bootstrap estimates and compared with the corresponding statistic obtained on the original sample. For example, the bootstrap parameter average value is:

$$\overline{\hat{\beta}}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b(\theta), \tag{4.26}$$

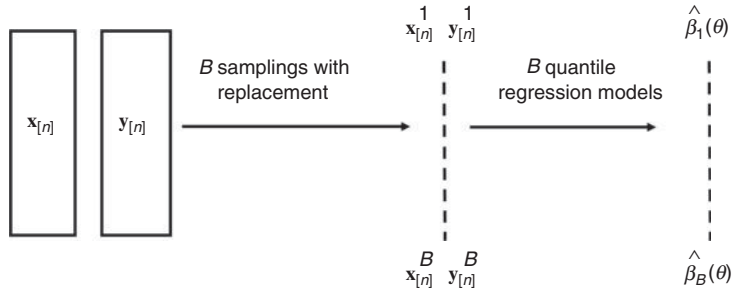

Figure 4.20 Representation of the bootstrap procedure with one quantile of interest and one explanatory variable. B *samples are drawn with replacement from the original dataset. Performing* B *QRs, a vector of bootstrap parameters is obtained for each quantile.*

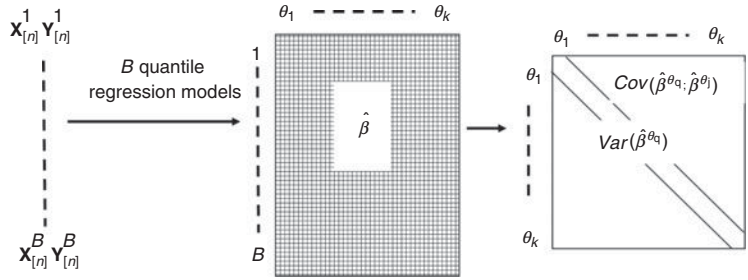

Figure 4.21 Representation of the bootstrap procedure in QR with k *quantiles of interest and one explanatory variable. Performing* B *QRs, a matrix of bootstrap parameters is obtained. From such a matrix, a variance-covariance matrix can be derived.*

and it can be compared with the value of $\hat{\beta}(\theta)$ obtained on the original sample. The standard error of the vector of parameter bootstrap estimates represents an estimate of the QR standard error useful in confidence intervals and hypothesis tests.

If k quantiles are considered, the bootstrap procedure produces a matrix of parameter estimates from which a variance-covariance matrix can be derived (Figure 4.21).

In cases of a multiple QR performed with p explanatory variables, the following bootstrap variance can be considered as an estimator of the asymptotic variance defined in Chapter 3 for each explanatory variable j and for each quantile q:

$$\hat{V}_{q,j} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_{b,j}(\theta_q) - \overline{\hat{\beta}}_j(\theta_q)\right)\left(\hat{\beta}_{b,j}(\theta_q) - \overline{\hat{\beta}}_j(\theta_q)\right)^{\top}, \tag{4.27}$$

where $j = 1, \ldots, p$; $q = 1, \ldots, k$ and $\overline{\hat{\beta}}_j(\theta_q) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{b,j}(\theta_q)$.

For each QR parameter, confidence intervals can be constructed using the estimated bootstrap standard errors in (4.27) or empirical quantiles of the vector of parameter estimates (percentile method).

In the first case, exploiting the asymptotic normal limit, a confidence interval for the generic jth parameter and the qth quantile, is:

$$\overline{\hat{\beta}}_j(\theta_q) \pm z_{\alpha/2} SD\left(\hat{\beta}_j(\theta_q)\right), \tag{4.28}$$

where $\overline{\hat{\beta}}_j(\theta_q)$ is the average value of the B bootstrap estimates, and $SD(\hat{\beta}_j(\theta_q))$ is the squared root of the variance in Equation (4.27).

The percentile method is based on the αth $(\hat{\beta}_j(\theta_q)_{lo})$ and $(1-\alpha)$th $(\hat{\beta}_j(\theta_q)_{up})$ percentiles of the cumulative distribution function of the bootstrap vector of parameter estimates:

$$\left[\hat{\beta}_j(\theta_q)_{lo}, \hat{\beta}_j(\theta_q)_{up}\right] = \left[\hat{F}(\alpha), \hat{F}(1-\alpha)\right], \tag{4.29}$$

where *lo* and *up* stand for, respectively, the lower and upper extreme of the confidence interval.

When the empirical distribution of the bootstrap estimates approximates a normal distribution, the two confidence intervals described before almost coincide. If the approximation is not reasonable, the percentile method should be preferred (Efron and Tibshirani 1986), even if this requires more replicates (Andrews and Buchinsky 2002). Moreover, Hahn (1995) proved that the bootstrap distribution obtained by the percentile method converges weakly to the appropriate limit distribution in probability, both for deterministic and random regressors.

To empirically show the effect of adopting a bootstrap procedure, the heterogeneous model $\mathbf{y}^{(6)}$ is considered. The x-coefficient of the median regression model is estimated through the xy-method with different numbers of replications (50, 100, 500, 1000). Figure 4.22 shows the QQ-plots of the estimates obtained in the four considered cases. The approximation to the normal distribution improves as the number of replications increases. Such improvement affects the differences between the asymptotic and percentile confidence intervals. In Figure 4.23, the distribution of the explanatory variable parameter vectors deriving from different B values is shown. Vertical lines represent the asymptotic confidence intervals obtained using a local estimate of the sparsity (solid line), the bootstrap confidence intervals (dashed lines) and the percentile confidence intervals (dotted lines). The percentile confidence intervals are always more precise than the others, and as the number of replications increases, the intervals become closer.

The capability of the bootstrap estimates to approximate a normal distribution also works in cases where the quantiles are different from the median, for example, in cases of the extreme quantile $\theta = 0.9$ (Figure 4.24).

One of the drawbacks of the bootstrap approach is related to the definition of the number of bootstrap replications (B). This biases the accuracy of the parameter estimation. A higher number of bootstrap replications will yield more accurate

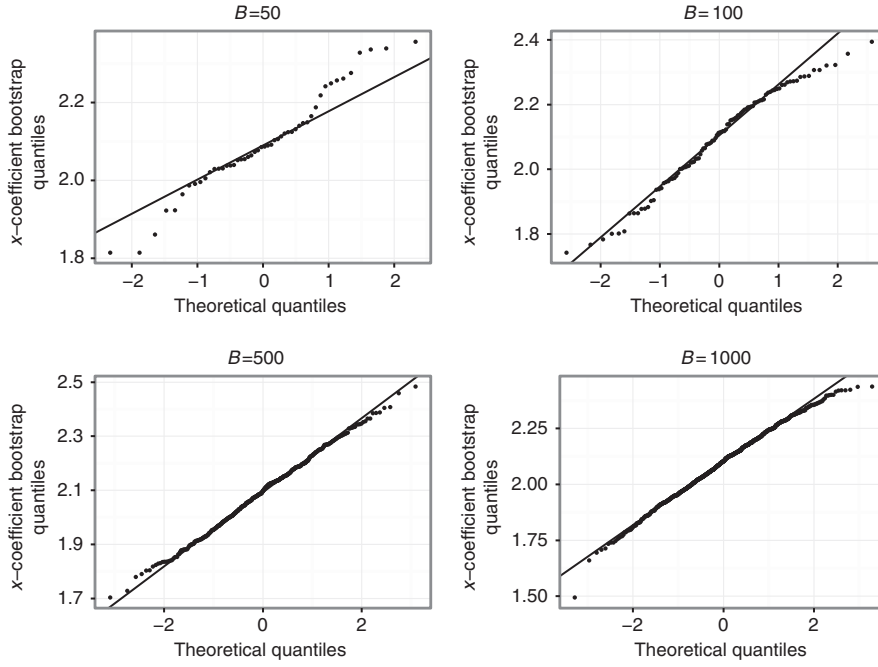

Figure 4.22 QQ-plots of the explanatory variable parameter vectors using different replications (heterogeneous model, θ = 0.5). The approximation to the normal distribution improves as the number of replications increases.

estimates of the parameter of interest. However, it is also important to take into account the computational cost related to a bootstrap procedure. The *xy*-pair method has been shown to provide accurate estimates, even in the presence of heteroskedasticity, but it requires the estimation of B QR models.

The bootstrap approach can also be used to estimate standard errors in cases of interquantile regression. As described in Chapter 3, the Student-*t* test permits to evaluate if the difference between the estimates of the same coefficient computed at different quantiles is significantly different from zero. The bootstrap standard errors can be obtained from the same procedure described in Figure 4.21. Once the bootstrap parameter matrix is obtained (B values of the β coefficient for each quantile), a parameter differences matrix can be derived by comparing pairs of bootstrap parameter vectors. From such a matrix, a variance-covariance matrix is derived containing the difference variances on the principal diagonal (Figure 4.25).

4.3.2.2 Method based on pivotal estimating functions

The bootstrap method based on pivotal estimating functions (Parzen *et al.* 1994) is known as the *pwy*-method. The name is derived from its originators: Parzen, Wei and

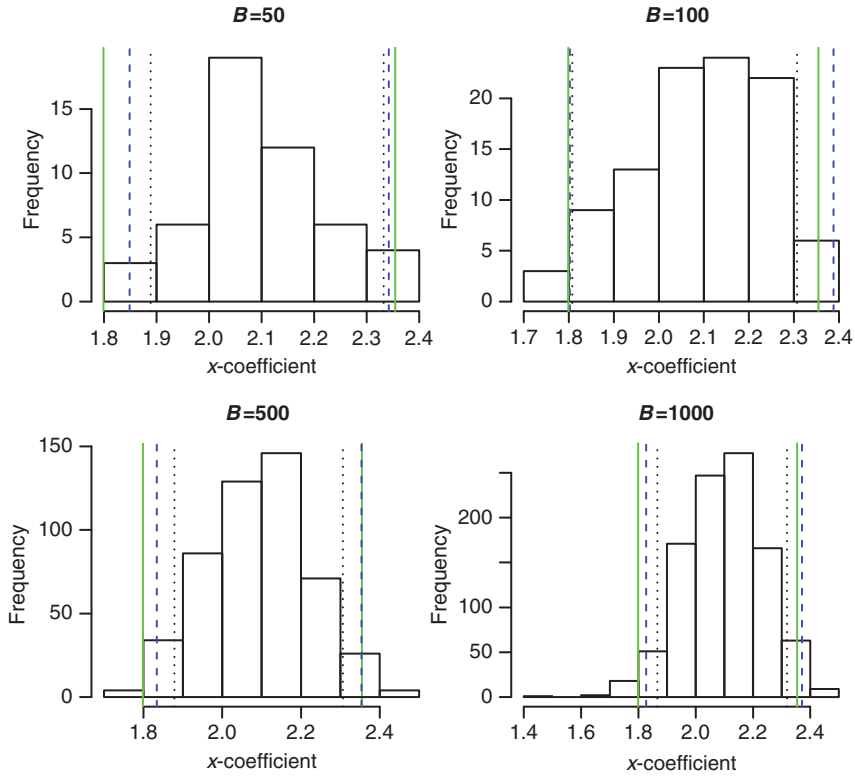

Figure 4.23 Histograms of the explanatory variable parameter vectors using different replications with asymptotic (solid lines), bootstrap (dashed lines) and percentile (dotted lines) confidence intervals (heterogeneous model). The length of the percentile confidence intervals is always shorter than the lengths derived from the other two methods. As the number of replications increases, the intervals become closer.

Ying. It requires resampling of the subgradient condition used in QR to obtain the parameter estimates:

$$S\left[\beta(\theta)\right] = \frac{1}{n}\sum_{i=1}^{n} \rho_\theta \left[y_i - \beta_0(\theta) - \beta_1(\theta)x_i\right]. \tag{4.30}$$

The originality of the approach is related to the authors' intuition that $S\left[\beta(\theta)\right]$ is a pivotal quantity for the true θ-th QR parameter. Thus, its distribution can be generated precisely by a random vector, U, representing a weighted sum of independent and centred Bernoulli variables. According to the *pwy*-method, the condition in Equation (4.30) is solved equal to u, a given realization of U, for a defined number of times (B replications). The resampling yields the empirical distribution of

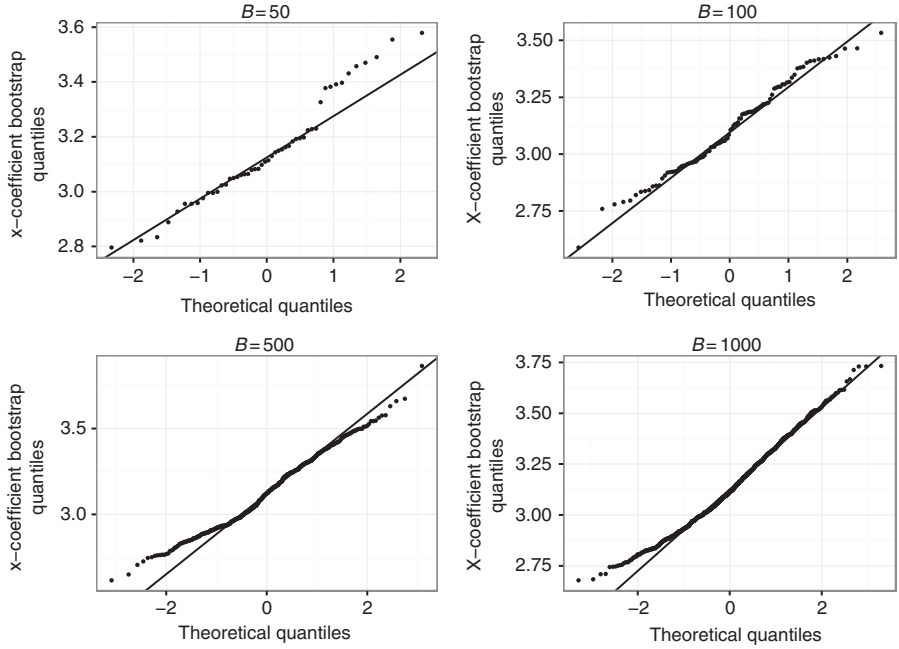

Figure 4.24 QQ-plots of the explanatory variable parameter vectors using different replications (heterogeneous model, θ = 0.9). The approximation to the normal distribution improves as the number of replications increases.

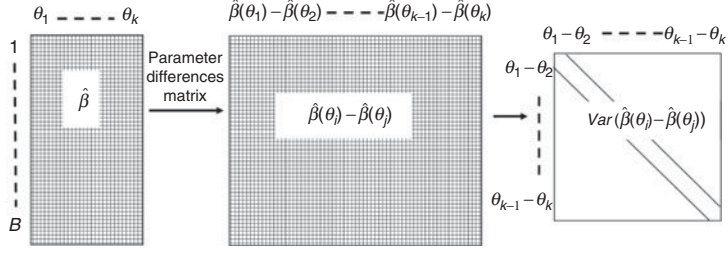

Figure 4.25 Representation of the bootstrap procedure in interquantile regression with k *quantiles of interest and one explanatory variable.*

β_U, which approximates to a corresponding theoretical distribution. Parzen, Wei and Ying demonstrated that the conditional distribution of $(\widetilde{\beta} - \beta_U)$ is asymptotically identical to the unconditional distribution of $(\widehat{\beta} - \beta_0)$ where $\widetilde{\beta}$ is the observed $\widehat{\beta}$.

Like any bootstrap method, the *pwy*-method requires an arbitrary choice of the number of replications. Moreover it is very time consuming because it is necessary to solve the gradient condition for each replication.

4.3.2.3 Markov chain marginal bootstrap

Markov chain marginal bootstrap (*mcmb*) aims to reduce computational costs deriving from the use of the previously described bootstrap procedures. The underlying idea is to solve one-dimensional equations for each bootstrap replication instead of a p-dimensional system required by the other bootstrap approaches, where p represents the number of explanatory variables. The approach is named Markov chain marginal bootstrap because, in common with Markov chain Monte Carlo methods, it aims to reduce a high dimensional problem into a lower-dimensional one. The final result of the method is a Markov chain for each component of the regression coefficient to be estimated. The approach was firstly proposed by He and Hu (2002) for constructing confidence intervals in a wide class of M-estimators and then generalized to QR by Kocherginsky (2003) and Kocherginsky *et al.* (2005).

The second moment of the obtained Markov chain, $n^{1/2}(\hat{\beta}^{(m)}(\theta) - \hat{\beta}(\theta))$ where $m = 1, \ldots, M$ and M is the length of the Markov chain, is able to approximate the coefficient variance-covariance matrix, even in cases of heteroskedasticity. Details of the algorithm can be found in He and Hu (2002), Kocherginsky (2003) and Kocherginsky *et al.* (2005).

In the following, the main features of the method are outlined:

- The asymptotic validity of *mcmb* requires that both the number of observations and the length of the Markov chain go to infinity. It follows that it may not be suitable for problems of small sample sizes.
- If the *mcmb* sequence is affected by autocorrelation, the accuracy in estimating the variance-covariance matrix is reduced. Kocherginsky (2003) tackled such a problem and solved it through a very simple transformation: the standardization of the explanatory variables (*mcmb-A* version of the method).
- As *mcmb-A* is valid only in cases of independent and identically distributed (i.i.d) errors, the transformation of the estimating equations (*mcmb-AB* version of the method) allows the method to be extended to include non-independent and identically distributed (ni.i.d.) error models.
- The surplus value of *mcmb* in reducing computer time increases with p.

4.3.2.4 Empirical comparisons among bootstrap methods in QR

In this section, the *xy*, *pwy* and *mcmb* methods are compared in terms of their capability to estimate parameter values and standard errors. Differences among the three bootstrap methods can be better highlighted by comparing results of

Table 4.9 Comparison of coefficient values with and without the use of bootstrap methods ($\theta = 0.5$). Bootstrap methods show different results particularly in the heterogeneous model.

	Homogeneous model		Heterogeneous model	
	Constant	Coefficient	Constant	Coefficient
QR	0.948	2.007	0.204	2.103
$xy_{B=200}$	0.970	2.005	0.367	2.088
$xy_{B=500}$	0.959	2.006	0.303	2.094
$xy_{B=1000}$	0.948	2.007	0.333	2.091
$pwy_{B=200}$	0.955	2.007	0.359	2.089
$pwy_{B=500}$	0.961	2.006	0.321	2.092
$pwy_{B=1000}$	0.954	2.007	0.304	2.094
$mcmb_{K=200}$	0.955	2.007	0.456	2.077
$mcmb_{K=500}$	0.955	2.006	0.362	2.087
$mcmb_{K=1000}$	0.958	2.006	0.384	2.084

the homogeneous and the heterogeneous model described at the beginning of Section 4.3.2.

In the first row of Table 4.9, the median regression parameter values in the homogeneous model (first two columns) and in the heterogeneous model (third and fourth columns) obtained for the whole sample are shown. In the other rows, the average estimates are computed using different numbers of replications for each bootstrap method.

Regarding the capability to estimate the parameter values, it is evident that they are almost the same in a homogeneous model. However, in a heterogeneous model the estimates of the *xy*-method are closer to the median regression values than the other methods, and the *mcmb* method produces the worst results.

The comparisons in terms of standard errors (Table 4.10) also take into account the values derived from some of the direct estimates methods described in Chapter 3: ni.i.d. refers to the use of the bandwidth of Hall and Sheather (1988) to compute the sparsity function, while i.i.d. refers to the estimate of the asymptotic covariance matrix as proposed by Koenker and Basset (1978). Differences only emerge with the heterogeneous method when there are a low number of replications, with the *pwy* bootstrap showing the best performance and the *mcmb* method showing the worst performance.

Comparisons shown in Table 4.9 and Table 4.10 are based on a single run of B or K replications for the *mcmb*-method. The results depend on the sample variability and the bootstrap variability described in the beginning of Section 4.3.2. A Monte Carlo study could be advisable to provide more robust comparisons among the bootstrap methods. Several interesting simulation studies can be found in Parzen *et al.* (1994), Kocherginsky *et al.* (2005), Koenker (2005), and Buchinsky (1995).

Table 4.10 Comparison of standard errors with and without the use of bootstrap methods ($\theta = 0.5$). Bootstrap methods show different results particularly in the heterogeneous model.

	Homogeneous model		Heterogeneous model	
	Constant	Coefficient	Constant	Coefficient
i.i.d.	0.131	0.013	1.433	0.143
ni.i.d.	0.128	0.013	1.358	0.138
$xy_{B=200}$	0.121	0.012	1.441	0.144
$xy_{B=500}$	0.132	0.013	1.405	0.142
$xy_{B=1000}$	0.131	0.013	1.370	0.138
$pwy_{B=200}$	0.125	0.013	1.375	0.140
$pwy_{B=500}$	0.132	0.013	1.353	0.138
$pwy_{B=1000}$	0.135	0.013	1.415	0.143
$mcmb_{K=200}$	0.125	0.012	1.427	0.142
$mcmb_{K=500}$	0.127	0.013	1.366	0.137
$mcmb_{K=1000}$	0.129	0.013	1.343	0.135

4.4 Summary of key points

- In QR, centring the explanatory variables does not change the direction and the magnitude of the coefficients.
- QR exhibits four equivariance properties.
- The construction of an estimate of the density function of the response variable can be realized using two different approaches: simulating different conditional densities of the response variable by conditioning on specific values of the explanatory variables and estimating the entire conditional distribution of the response by exploiting the entire quantile process.
- QR allows the distribution of the response variable to be reconstructed almost perfectly.
- In the QR framework, resampling methods estimate parameter standard errors without requiring any assumption on the error distribution.

References

Abramson IS 1982 On bandwidth variation in kernel estimates-A square root law. *Annals of Statistics* **10**(4), 1217–1223.

Ahamada I and Flachaire E 2011 *Non-Parametric Econometrics*. Oxford University Press.

Andrews DWK and Buchinsky M 2002 On the number of bootstrap repetitions for BCa confidence intervals. *Econometric Theory* **18**, 962–984.

Box GEP and Cox DR 1964 An analysis of transformations (with Discussion). *Journal of the Royal Statistical Society: Series B* **26**, 211–252.

Buchinsky M 1995 Estimating the asymptotic covariance matrix for quantile regression models. A Monte Carlo study. *Journal of Econometrics* **68**, 303–338.

Buchinsky M 1998 Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources* **33**, 88–126.

Cade BS and Noon BR (2003) A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1**(8), 412–420.

Chen C 2005 Growth charts of body mass index (BMI) with quantile regression. *Proceedings of International Conference on Algorithmic Mathematics and Computer Science*, Las Vegas.

Chiandotto B and Gola MM 1999 Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti. *Rapporto finale del gruppo di Ricerca, Comitato nazionale per la valutazione del sistema universitario*. http://www.cnvsu.it/_library/downloadfile.asp?id=10717.

Davino C and Vistocco D 2008a The evaluation of university educational processes: a quantile regression approach. *Statistica* **3**, 281–292.

Davino C and Vistocco D 2008*b* Quantile regression for the evaluation of student satisfaction. *Italian Journal of Applied Statistics* **20**, 179–196.

Efron B and Tibshirani RJ 1986 Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**(1), 54–77.

Efron B and Tibshirani RJ 1998 *Introduction to the Bootstrap*. Chapman & Hall/CRC.

Gould W 1992 Quantile regression with bootstrapped standard errors. *STATA Technical Bulletin*, 19–21.

Gujarati DN 2003 *Basic Econometrics*. International Edition. McGrawHill.

Hahn J 1995 Bootstrapping quantile regression estimators. *Econometric Theory* **11**, 105–121.

Hall P and Sheather S 1988 On the distribution of a studentized quantile. *Journal of the Royal Statistical Society: Series* B **50**, 381–391.

Hao L and Naiman D Q 2007 *Quantile Regression*. SAGE Publications, Inc.

He X and Hu F 2002 Markov Chain Marginal Bootstrap. *Journal of the American Statistical Association* **97**(459), 783–795.

Hyndman RJ, Bashtannyk DM and Grunwald GK 1996 Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* **5**(4), 315–336.

Kocherginsky M 2003 *Extensions of the Markov Chain Marginal Bootstrap*. PhD Thesis, University of Illinois Urbana-Champaign.

Kocherginsky M, He X and Mu Y 2005 Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics* **14**(1), 41–55.

Koenker R 2005 *Quantile Regression*. Cambridge University Press.

Koenker RW and Basset G 1978 Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker RW and Basset G 1982 Robust tests for heteroskedasticity based on regression quantiles. *Econometrica* **50**(1), 43–61.

Koenker R. and Machado J 1999 Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94**, 1296–1310.

Koenker R and Xiao Z 2002 Inference on the quantile regression process. *Econometrica* **70**(4), 1583–1612.

Kuan C-M 2007 *An Introduction to Quantile Regression*. Institute of Economics, Academia Sinica.

Manning WG 1998 The logged dependent variable, heteroskedasticity, and the retransformation problem. *Journal of Health Economics* **17**, 283–295.

Parzen MI, Wei LJ and Ying Z 1994 A resampling method based on pivotal estimating functions. *Biometrika* **81**(2), 341–50.

Santamaria F, Montella S, Greco L, Valerio G, Franzese A, Maniscalco M, Fiorentino G, Peroni D, Pietrobelli A, De Stefano S, Sperl F. and Boner AL 2011 Obesity duration is associated to pulmonary function impairment in obese subjects. *Obesity* (Silver Spring) **19**(8), 1623–1628.

Shapiro SS and Wilk MB 1965 An analysis of variance test for normality (complete sample). *Biometrika* **52**, 591–611.

Shapiro SS and Wilk MB 1968 Approximations for the null distribution of the W statistic. *Technometrics* **10**, 861–866.

Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

White H 1980 A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**(4), 817–838.

Van Kerm P 2003 Adaptive kernel density estimation. *The Stata Journal* **3**(2), 148–156.

von Eye A and Schuster C 1998 *Regression Analysis for Social Sciences*. Academic Press, San Diego.

5

Models with dependent and with non-identically distributed data

Introduction

This chapter focuses on the quantile regression estimators for models characterized by heteroskedastic and by dependent errors. Section 5.1 considers the precision of the quantile regression model in the case of independent and identically distributed errors, taking a closer look at the computation of confidence intervals and hypothesis testing on each estimated coefficient. The empirical example further analyzes the wage equation introduced in Chapter 3. Section 5.2 extends the analysis to the case of non-identically distributed errors, discussing different ways to verify the presence of heteroskedasticity in the data. The examples considers the series of changes in consumption analyzed in Section 3.1.2 and the Italian GDP growth rate. Section 5.3 takes into account the case of dependent observations and discusses the estimation of an exchange rate equation characterized by serially correlated errors.

5.1 A closer look at the scale parameter, the independent and identically distributed case

5.1.1 Estimating the variance of quantile regressions

Consider the linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ with $p = 2$ unknown coefficients, independent and identically distributed (i.i.d.) errors, in a sample of

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

size n. The i.i.d. errors have common positive density f at the selected quantile, $f(F^{-1}(\theta)) > 0$. The explanatory variables are comprised in the $(n, 2)$ matrix $\mathbf{X} = [\mathbf{1}\ \mathbf{x}]$ where the terms $\mathbf{1}$ and $\mathbf{x}$ represent $(n, 1)$ column vectors; the quadratic form of the $\mathbf{X}$ matrix is assumed to converge to the positive definite matrix $\mathbf{D} = lim_{n\to\infty} 1/n\mathbf{X}^\top\mathbf{X}$. Then in the i.i.d. case the asymptotic distribution of $\widehat{\boldsymbol{\beta}}(\theta)$, the quantile regression (QR) estimator of the coefficient vector $\boldsymbol{\beta}(\theta)^\top = [\beta_0\ \beta_1]$, is given by:

$$\sqrt{n}[\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)] \to N(0, \omega^2(\theta)\mathbf{D}^{-1}), \tag{5.1}$$

where the scale parameter of the model $\omega^2(\theta)$, at the selected quantile θ, is defined as:

$$\omega^2(\theta) = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2}. \tag{5.2}$$

Equation (5.2) states that the precision of the estimated coefficients is inversely related to the density at the selected quantile, $f(F^{-1}(\theta))$. High density involves little dispersion and high precision, while low density involves great dispersion and low precision.

The term $f(F^{-1}(\theta))$ is unknown and has to be estimated. Many different estimators have been proposed. Siddiqui (1960) considers the following:

$$s(t) = \frac{1}{f(F^{-1}(\theta))} = \frac{F^{-1}(t+\mathrm{h}) - F^{-1}(t-\mathrm{h})}{2\mathrm{h}}, \tag{5.3}$$

that has a nice geometric interpretation. The reciprocal of the density at a given quantile, the so-called sparsity function $s(t) = 1/f(F^{-1}(\theta))$, can be computed by differentiating the quantile function $F^{-1}(\theta)$, $s(t) = 1/f(F^{-1}(\theta)) = dF^{-1}(t)/dt$, and thus it represents the slope of the tangent to the quantile function at point t. This slope can be approximated by the slope of the secant to the quantile function at points $t+\mathrm{h}$ and $t-\mathrm{h}$, as stated in the right-hand side term of Equation (5.3)

Besides the nice geometric interpretation, depicted in Figure 5.1, the bandwidth h and the F function have to be defined. Referring to the standard normal distribution Φ, Koenker (2005) suggests to select the bandwidth:

$$\mathrm{h} = n^{-1/5}\left[\frac{4.5\phi^4(\Phi^{-1}(t))}{(2\Phi^{-1}(t)^2+1)^2}\right]^{1/5}, \tag{5.4}$$

while Koenker and Machado (1999) consider the bandwidth:

$$\mathrm{h} = n^{-1/3} z_\alpha^{2/3}\left[\frac{1.5\phi^4(\Phi^{-1}(t))}{(2\Phi^{-1}(t)^2+1)^2}\right]^{1/3}, \tag{5.5}$$

where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha/2$.

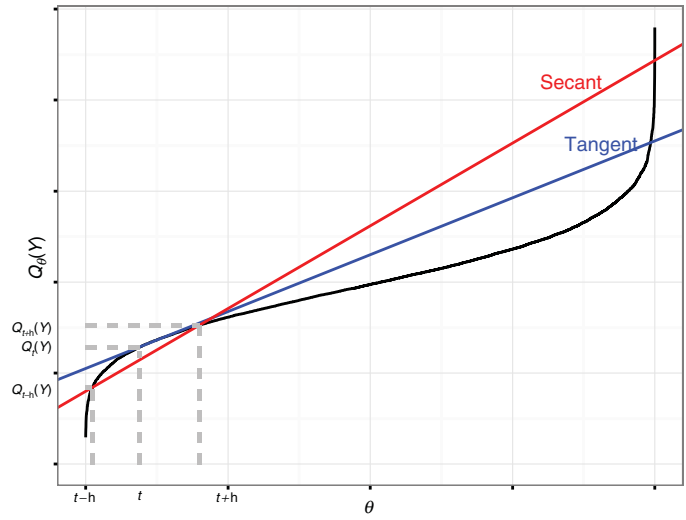

Figure 5.1 The sparsity function coincides with the slope of the tangent at the selected quantile. It can be approximated by the slope of the secant, given by the difference between the two points where the secant intersects the quantile function, divided by the width of the interval 2h, as reported in Equation (5.3).

Finally, in Equation (5.3) the quantile function $F^{-1}(t \pm \mathrm{h})$ has to be estimated. The empirical quantile function can be used for the purpose, and an estimate of $F^{-1}(t)$ is provided by the residual of the QR, $y_i - \hat{\beta}_0(t) - \hat{\beta}_1(t)x_i = u_i$, where $\hat{\boldsymbol{\beta}}^\top(t) = \left[\hat{\beta}_0(t) \quad \hat{\beta}_1(t)\right]$ are the QR estimates of the true coefficients $\boldsymbol{\beta}(\theta)$ at the quantile θ. Alternatively, a different estimate of the quantile function $F^{-1}(t)$ considers the sample mean of x_i to compute pseudo-fitted values of Y, given by $\hat{\beta}_0(t) + \hat{\beta}_1(t)\bar{x}$.

A completely different approach, easier but computationally intensive, is also available. Instead of estimating the density function, a bootstrap approach can be implemented. The most successful approach is the design matrix bootstrap (Buchinsky 1995) which instead of focusing on the dispersion as measured by the sparsity function and by the nuisance parameter $\omega^2(\theta)$, focuses on the variability of the estimated coefficients. A number of π subsamples, each of size m, are drawn from the empirical joint distribution of X and Y, and the QR coefficients are estimated in each subsample. The variability of the π estimates with respect to their average provides an estimate of the variance of the QR estimated coefficients. In addition, the ordered estimated values of each coefficient allow to easily compute confidence intervals. The lower and upper bounds are given by the $B(\alpha/2)$ element and the $B(1-\alpha/2)$ element of the ordered estimated coefficients. This approach has the advantage of being very reliable even in case of non-identically distributed errors. For further details, see Chapter 4, Section 4.3.2.

5.1.2 Confidence intervals and hypothesis testing on the estimated coefficients

Once the sparsity function has been computed, Equation (5.2) defines the nuisance parameter of the QR, while the estimator of the variance covariance matrix of the regression coefficients is:

$$E[\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)][\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)]^{\top} = \omega^2(\theta)\mathbf{D}^{-1}. \tag{5.6}$$

The square root of the diagonal elements of the estimated matrix in (5.6) provides the standard errors of the estimated regression coefficients, $se(\widehat{\beta}_{\mathcal{P}}(\theta))$, $\mathcal{P} = 1, \ldots, p$. The standard errors allow to compute confidence intervals and to test the hypothesis H_0: $\beta_p(\theta) = 0$. In particular the confidence interval is given by:

$$P(\widehat{\beta}_{\mathcal{P}}(\theta) - se(\widehat{\beta}_{\mathcal{P}}(\theta)) * z_{1-\alpha/2} \leq \beta_{\mathcal{P}}(\theta) \leq \widehat{\beta}_{\mathcal{P}}(\theta) + se(\widehat{\beta}_{\mathcal{P}}(\theta)) * z_{1-\alpha/2}) = 1 - \alpha, \tag{5.7}$$

and the Student-t test with $n - p$ degrees of freedom to verify the above null hypothesis is $t = \widehat{\beta}_{\mathcal{P}}(\theta)/se(\widehat{\beta}_{\mathcal{P}}(\theta))$.

5.1.3 Example for the i.i.d. case

The wage equation of Chapter 3 is defined as:

$$lwage_i = \beta_0 + \beta_1 education_i + \beta_2\, age_i + e_i.$$

At the median, the vector of estimated coefficients is $\widehat{\boldsymbol{\beta}}(\theta)^{\top} = \begin{bmatrix} 0.326 & 0.064 & 0.027 \end{bmatrix}$.

An estimate of the density is given by $\widehat{f(F^{-1}(\theta)} = \widehat{f(F(0.50)^{-1}} = 0.9397$, and the sparsity function is estimated as $s(\theta) = 1/\widehat{f(F(\theta)^{-1}} = 1/0.9397 = 1.064$. Figure 5.2 displays the density and quantile function of the residuals at the median regression of the above wage equation. It can be seen that, in Figure 5.2(b), at the quantile $\theta = 0.50$ the residuals are equal or close to zero. Figure 5.3 presents the density and quantile function of the residuals at the first quartile, and in Figure 5.3(b) it can be seen that at $\theta = 0.25$ the residuals are equal or close to zero. The estimated value of $f(F(\theta)^{-1})$ at the first quartile is $\widehat{f(F(0.25)^{-1}} = 0.6879$. In general, for any estimated QR, $f(F(\theta)^{-1})$ is computed by the height of the histogram at that particular quantile, for those residuals at and around zero.

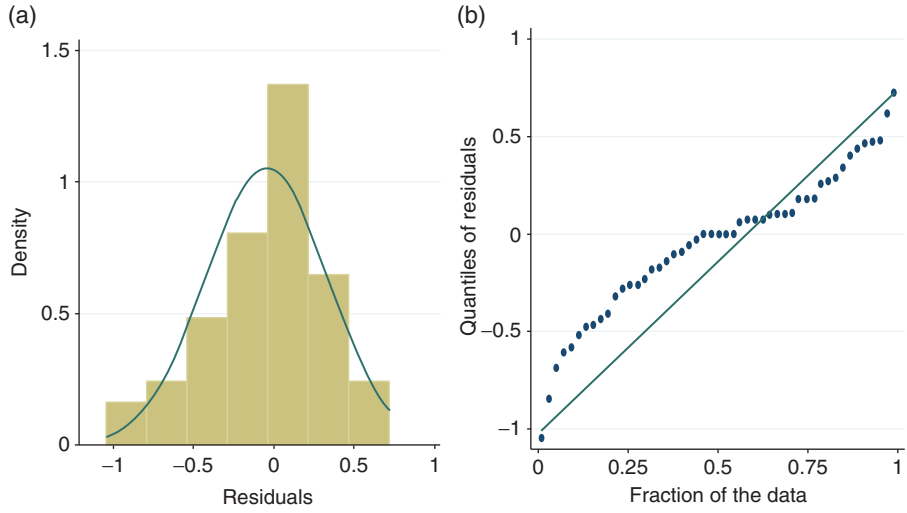

Figure 5.2 (a) Density and (b) quantile function of the residuals at the median regression of the wage equation. In (b) the residuals equal or close to zero are around the median, at $\theta = 0.50$.

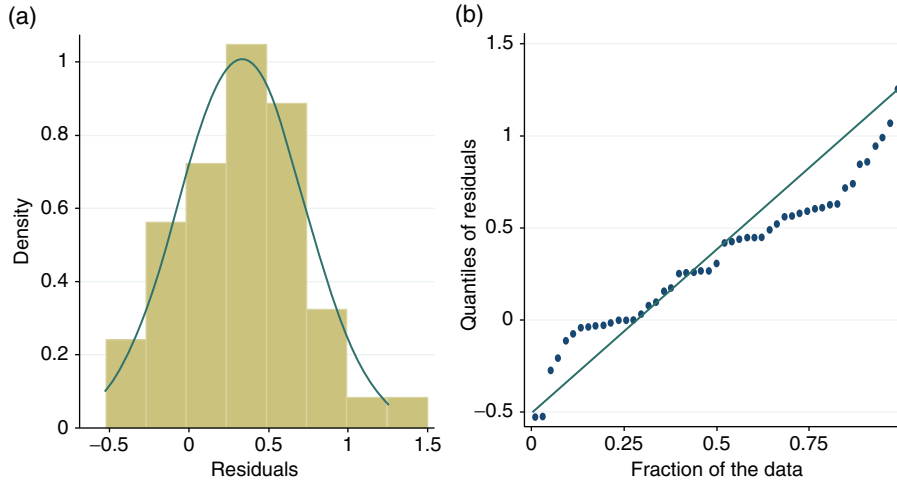

Figure 5.3 (a) Density and (b) quantile function of the residuals at the first quartile regression of the wage equation. In (b) the residuals equal or close to zero are around the first quartile, at $\theta = 0.25$.

Having an estimate of the density at the median, the scale parameter is estimated as $\widehat{\omega}^2(\theta) = \frac{\theta(1-\theta)}{\widehat{f(F^{-1}(\theta))^2}} = \frac{0.25}{0.9397^2} = 0.283$ and the estimated variance-covariance matrix of the regression coefficients is:

$$\widehat{\omega}^2(\theta)\mathbf{D}^{-1} = \widehat{\omega}^2(\theta)(\mathbf{X}^\top\mathbf{X})^{-1} =$$

$$= 0.283 \begin{bmatrix} 49 & 712 & 1961 \\ 712 & 11056 & 28798 \\ 1961 & 28798 & 82671 \end{bmatrix}^{-1}$$

$$= 0.283 \begin{bmatrix} 0.599 & -0.017 & -0.0083 \\ -0.017 & 0.0014 & -0.0001 \\ -0.0083 & -0.0001 & 0.00024 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.169 & -0.0048 & -0.0023 \\ -0.0048 & 0.0004 & -0.00002 \\ -0.0023 & -0.00002 & 0.00006 \end{bmatrix}.$$

The square root of the diagonal elements of the final matrix provide the standard errors of the estimated coefficients, which measure their precision. For the intercept the estimated standard error is $se(\widehat{\beta_0(\theta)}) = \sqrt{0.169} = 0.411$, for the education coefficient $se(\widehat{\beta_1(\theta)}) = \sqrt{0.0004} = 0.02$ and for the age coefficient $se(\widehat{\beta_2(\theta)}) = \sqrt{0.00006} = 0.008$. These values are quite similar to the standard errors reported in Table 3.8. It is now easy to compute confidence intervals:

$$P(\widehat{\beta}_0(\theta) - se(\widehat{\beta_0(\theta)}) * z_{1-\alpha/2} \leq \beta_0(\theta) \leq \widehat{\beta}_0(\theta) + se(\widehat{\beta_0(\theta)}) * z_{1-\alpha/2}) =$$
$$= P(0.326 - 0.411 * z_{1-\alpha/2} \leq \beta_0(\theta) \leq 0.326 + 0.411 * z_{1-\alpha/2}) = 1 - \alpha;$$

$$P(\widehat{\beta}_1(\theta) - se(\widehat{\beta_1(\theta)}) * z_{1-\alpha/2} \leq \beta_1(\theta) \leq \widehat{\beta}_1(\theta) + se(\widehat{\beta_1(\theta)}) * z_{1-\alpha/2}) =$$
$$= P(0.064 - 0.02 * z_{1-\alpha/2} \leq \beta_1(\theta) \leq 0.064 + 0.02 * z_{1-\alpha/2}) = 1 - \alpha;$$

$$P(\widehat{\beta}_2(\theta) - se(\widehat{\beta_2(\theta)}) * z_{1-\alpha/2} \leq \beta_2(\theta) \leq \widehat{\beta}_2(\theta) + se(\widehat{\beta_2(\theta)}) * z_{1-\alpha/2}) =$$
$$= P(0.027 - 0.008 * z_{1-\alpha/2} \leq \beta_2(\theta) \leq 0.027 + 0.008 * z_{1-\alpha/2}) = 1 - \alpha.$$

At the median regression, by choosing $\alpha = 5\%$, the above confidence intervals become

$$P(-0.479 \leq \beta_0(\theta) \leq 1.131) = 95\%;$$
$$P(0.025 \leq \beta_1(\theta) \leq 0.103) = 95\%;$$
$$P(0.011 \leq \beta_2(\theta) \leq 0.042) = 95\%.$$

Tests on the individual regression coefficients are given by the ratio between the estimated coefficient and its estimated standard error, and are asymptotically distributed as Student-t with $n - p$ degrees of freedom. Under the null H_0:$\beta_{\mathcal{P}}(\theta) = 0$, $\mathcal{P} = 1, \ldots, p$; $t(\beta_0(\theta)) = 0.326/0.411 = 0.793$; $t(\beta_1(\theta)) = 0.064/0.02 = 3.2$; $t(\beta_2(\theta)) = 0.027/0.008 = 3.375$. The null is rejected for both education and age coefficients while it is not rejected for the intercept, just as reported in Table 3.8.

Next, the variability of the model and the confidence intervals can be computed implementing the bootstrap approach. In this approach 1000 samples of size $n = 49$ are randomly drawn with replacement from the wage dataset. In each of the 1000 samples the coefficients of the wage equation are estimated at the median regression. The dispersion of the 1000 estimates for each coefficient is computed to provide the estimated variance of each regression coefficient and the corresponding confidence interval. The estimated standard errors are $se(\widehat{\beta_0(\theta)}) = 0.364$; $se(\widehat{\beta_1(\theta)}) = 0.021$; $se(\widehat{\beta_2(\theta)}) = 0.008$. These values do not sizeably differ from the previous results and yield the following 95% confidence intervals, computed implementing the percentile method:

$$P(-0.401 \leq \beta_0(0.5) \leq 1.08) = 95\%;$$

$$P(0.021 \leq \beta_1(0.5) \leq 0.103) = 95\%;$$

$$P(0.008 \leq \beta_2(0.5) \leq 0.045) = 95\%.$$

The latter do not differ much from the confidence intervals previously computed.

5.2 The non-identically distributed case

Real data do not generally abide by the i.i.d. assumption. There are two ways to diverge from this assumption, either the errors are not independent or they are not identically distributed, and of course they can also be both dependent and non-identically distributed. The errors are not identically distributed when their variances are not constant but do change across the sample. As mentioned in Chapter 3, a typical example is given by consumption as a function of income. At low income levels the greatest fraction of income is assigned to consumption, but as income increases there is a wider choice for individuals between saving and spending. This implies that, by aggregating the consumption decisions of many individuals, at the higher incomes the aggregate consumption is characterized by a greater variability with respect to the consumption decisions of the group of individuals earning lower incomes. The effect is a variability in consumption which increases with income: σ_i changes across the sample, and the explanatory variable x_i, that is the income at the i-th observation, can be considered the source of heteroskedasticity, $\sigma_i = h(x_i)$.

Heteroskedasticity generally occurs in cross-sectional data, however it can occur in time series and in all sorts of data. The heteroskedastic regression model can be written as $y_i = \beta_0 + \beta_1 x_i + \sigma_i e_i = \beta_0 + \beta_1 x_i + u_i$ with $e_i \sim$ i.i.d., and the presence of heteroskedasticity modifies the variance-covariance matrix of the QR coefficients as follows:

$$E[\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)][\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)]^\top = \mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1}, \tag{5.8}$$

where the matrices $\mathbf{D}$ and $\mathbf{D}_1(\theta)$ are defined as $\mathbf{D} = lim\mathbf{X}^\top\mathbf{X}, \mathbf{D}_1(\theta) = lim\mathbf{X}^\top \boldsymbol{\Gamma}^{-1}\mathbf{X}$, and $\boldsymbol{\Gamma} = diag(\omega_i^2(\theta))$ where $\omega_i^2(\theta) = \frac{\theta(1-\theta)}{f_i(F^{-1}(\theta))^2}$.

The changing density of u_i in the Γ matrix, $f_i(F^{-1}(\theta))$, can be split into a constant part given by the density of e_i, $f(F^{-1}(\theta))$, times a changing one given by $\sigma_i = h(x_i)$. In the i.i.d. case the matrix $\mathbf{D}_1(\theta)$ simplifies into $\mathbf{D}$ thus Equation (5.8) coincides with Equation (5.1). Indeed $f_i(F^{-1}(\theta)) = f(F^{-1}(\theta))$ is constant in the case of i.i.d. errors and $\Gamma = diag(\omega^2(\theta)) = \omega^2(\theta)\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, $\mathbf{D}_1(\theta) = \mathbf{D}$, and in Equation (5.8) $\mathbf{D}$ and $\mathbf{D}_1(\theta)^{-1}$ cancel.

Ignoring heteroskedasticity, in ordinary least square (OLS) as in the QR framework, leads to faulty inference. Indeed a single estimate of $\omega(\theta)$ is used instead of the changing estimates $\omega_i(\theta)$ that characterize the dataset. It is like using an average value instead of the appropriate different values to implement inference.

A test for heteroskedasticity in the QR framework looks at the coefficients estimated at different quantiles. Indeed the changing σ_i have an impact on the estimated coefficients of the QRs, and cause $\boldsymbol{\beta}(\theta)$ to change across quantiles. For instance, in the consumption example of Section 3.1.2 and Figure 3.10, where the dispersion increases over time, the fitted line at the 90th quantile will have a greater slope than the fitted line at the 10th quantile, in order to appropriately partition the observations – and the residuals – in two groups, 90% below and 10% above the fitted line. Analogously, the 10th quantile has a much smaller slope in order to partition 10% of the residuals below and 90% above the fitted line. Figure 5.4 presents the data on changes in consumption over time and the estimated QRs. The fitted lines are very

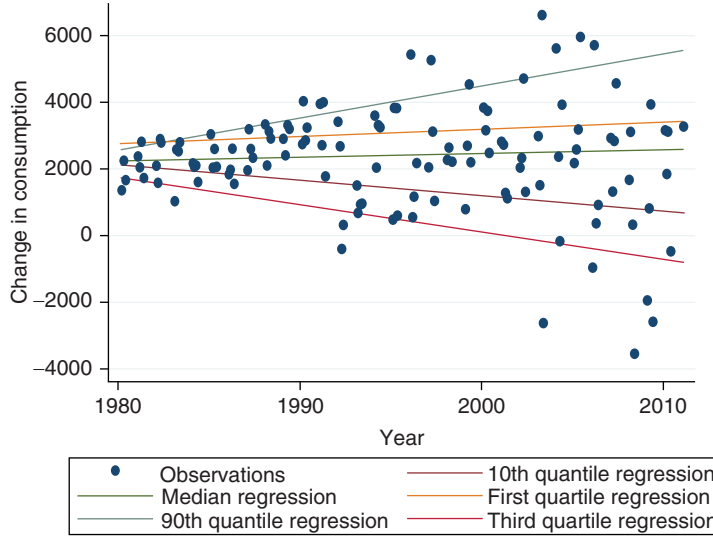

Figure 5.4 QRs in the case of heteroskedastic observations. Italian quarterly data of changes in consumption from 1980 to 2011, in a sample of size n = *124. The variance at the beginning of the sample is sizeably smaller than at the end of the sample. Data in the 1980s are realizations of a density* f_{i} *characterized by a smaller variance than the density* f_{j} *generating the observations in the last decade.*

Table 5.1 QR estimates, change in consumption dataset.

θ-quantile	0.10	0.25	0.50	0.75	0.90
Time	−8.176	−4.555	1.212	1.960	9.611
Student-t	(1.30)	(2.51)	(0.74)	(1.32)	(3.56)

Absolute value of Student-t statistics in parentheses, sample size $n = 124$.

close by at the beginning of the sample but diverge at the end of the sample, thus assuming a fan shaped pattern in order to accommodate the wider dispersion of the data occurring at the end of the sample. Table 5.1 reports the QR estimates of the slope coefficient at the selected quantiles.

In general, while in the i.i.d. case the estimated regressions are parallel lines differing from one another only in the intercept, with heteroskedasticity both the intercept and the slope coefficients change with the quantile. This leads to interpret the QR model as a random coefficient model, where the vector of coefficients can be decomposed in a fixed plus a random component, the latter depending on the quantile function $F^{-1}(\theta)$:

$$\beta(\theta) = \beta + \eta F^{-1}(\theta). \tag{5.9}$$

In the i.i.d. case only the intercept is variable and changes with the quantile, while all the other coefficients remain constant across quantiles and yield parallel regression quantile lines. Equation (5.9) in the i.i.d. case is $\beta_0(\theta) = \beta_0 + \eta_0 F^{-1}(\theta)$ for the intercept and $\beta_1(\theta) = \beta_1$ for the slope coefficient. For instance in Table 3.3 while the slope at the median, at the third quartile and at the 90th quantile is always equal to $\beta_1(\theta) = \beta_1 = 0.09$, the intercept assumes values $\beta_0(0.50) = 1.107, \beta_0(0.75) = 1.299$ and $\beta_0(0.90) = 1.505$, which can be interpreted as a constant plus a changing component intercept, with the changing component depending upon the selected quantile.

In the non-identically distributed case, instead, the random component affects both intercept and slope, and the latter is given by $\beta_1(\theta) = \beta_1 + \eta_1 F^{-1}(\theta)$. For instance the consumption model of Figure 5.4, where changes in consumption are a function of a time trend, provides marked differences in the estimated slopes, as can be seen in Table 5.1 where $\beta_1(0.50) = 1.212, \beta_1(0.75) = 1.960$ and $\beta_1(0.90) = 9.611$.

This is why a formal test for heteroskedasticity verifies if the differences in the estimated slopes across quantiles are statistically significant or if these differences are irrelevant and are simply due to a sample variation. Setting aside the intercept, which changes in both the i.i.d. and the non-identically distributed (i.i.d) case, the test considers the pairwise comparison between the same slope coefficient estimated at the i-th and the jth quantiles. For instance, in the comparison of one slope

coefficient estimated at two different quantiles, the null H_0: $\beta(\theta_i) = \beta(\theta_j)$ is verified by the test function

$$T = \frac{[\widehat{\beta}(\theta_i) - \widehat{\beta}(\theta_j)]^2}{var[\widehat{\beta}(\theta_i) - \widehat{\beta}(\theta_j)]},$$

where $var[\beta(\theta_i) - \beta(\theta_j)]$ is estimated by:

$$\begin{aligned} var[\beta(\theta_i) - \beta(\theta_j)] &= var\left[\beta(\theta_i)\right] + var\left[\beta(\theta_j)\right] - 2cov\left[\beta(\theta_i)\beta(\theta_j)\right] \\ &= \left[\omega^2(\theta_i) + \omega^2(\theta_j) - 2\omega(\theta_i\theta_j)\right] / n\ var(x) = \\ &= \frac{\left[\frac{\theta_i(1-\theta_i)}{f(F^{-1}(\theta_i))^2} + \frac{\theta_j(1-\theta_j)}{f(F^{-1}(\theta_j))^2} - 2\frac{\min(\theta_i,\theta_j)-\theta_i\theta_j}{f(F^{-1}(\theta_i))f(F^{-1}(\theta_j))}\right]}{(n\ var(x))}, \end{aligned} \tag{5.10}$$

where $\beta(\theta_i)$ is the estimated slope at the i-th QR and $\beta(\theta_j)$ its analog at the jth quantile. Equation (5.10) comprises the variability at θ_i, as measured by the scale parameter $\omega^2(\theta_i)$, the variability at θ_j, as measured by the scale parameter $\omega^2(\theta_j)$, plus the covariance between the i-th and the jth quantiles, provided by $\omega(\theta_i\theta_j)$. In the more general case of pairwise comparisons of $k > 2$ QRs, the test function is given by:

$$T = n[\beta(\theta_i) - \beta(\theta_j)]^\top [\mathbf{\Delta\Omega\Delta}^\top \otimes \mathbf{\Psi D}^{-1}\mathbf{\Psi}^\top]^{-1} \left[\beta(\theta_i) - \beta(\theta_j)\right], \tag{5.11}$$

where the matrix $\Omega = \left[\omega(\theta_i\theta_j)\right]$ has typical element:

$$\omega(\theta_i\theta_j) = \frac{\min(\theta_i, \theta_j) - \theta_i\theta_j}{f(F^{-1}(\theta_i))f(F^{-1}(\theta_j))},$$

which for $i = j$ yields $\omega^2(\theta_i)$ while for $i \neq j$ provides the covariances across different quantiles $\omega(\theta_i\theta_j)$; $\mathbf{\Psi}$ and $\mathbf{\Delta}$ are selection matrices to locate the relevant elements of $\mathbf{D}^{-1}$ and $\mathbf{\Omega}$. In the previous simplest case, which compares one coefficient estimated at two different quantiles, the selection terms to pin down the appropriate elements of the covariance matrix $\mathbf{\Omega}$, namely $\omega^2(\theta_i)$, $\omega^2(\theta_j)$, $\omega(\theta_i\theta_j)$, and of the quadratic form of the explanatory variables $\mathbf{D} = \mathbf{X}^\top\mathbf{X}$, namely $n\ var(X)$, are, respectively, the vectors $\Delta = [\ 1\ \ -1\ ]$ and $\mathbf{\Psi} = [\ 0\ \ 1\ ]$.

In the general case of $k > 2$, the $(k-1, 1)$ vector defining the comparisons between estimates computed at different QRs, together with the selection matrices are defined as:

$$[\beta(\theta_i) - \beta(\theta_j)] = \begin{bmatrix} \beta(\theta_j) - \beta(\theta_i) \\ \beta(\theta_{j+1}) - \beta(\theta_j) \\ \beta(\theta_{j+2}) - \beta(\theta_{j+1}) \end{bmatrix}, \mathbf{\Delta} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

$$\mathbf{\Psi} = \begin{bmatrix} 0 & \vdots & \mathbf{I}_{p-1} \end{bmatrix},$$

where $\mathbf{I}_{[p-1]}$ is the $(p-1)$ identity matrix.[1] Under the null of homoskedasticity, H_0: $\beta(\theta_i) = \beta(\theta_j)$, Equation (5.11) is asymptotically distributed as a χ^2 with degree of freedom equal to $(k-1)(p-1)$, where k is the number of QRs implemented, $k-1$ the number of comparisons under test, p the number of parameters in the equation and $p-1$ the number of slope parameters excluding the intercept (Koenker and Bassett 1982).

5.2.1 Example for the non-identically distributed case

An example for the non-identically distributed data is provided by the growth rate of the Italian GDP from 1971 to 2009, defined as $\text{GDP}_i = \frac{(gdp_i - gdp_{i-1})}{gdp_{i-1}}$. Figure 5.5 presents the scatter plot of the GDP rate of growth over time.[2] The model explains the GDP growth rate as a function of a time trend and a constant term, $\text{GDP}_i = \beta_0 + \beta_1 year_i + u_i$. The results of the estimated median regression are in Table 5.2, which reports a negative slope coefficient thus signaling a decreasing growth rate over time. Figure 5.5 presents also the OLS and the median regression fitted lines, which are very close to one another, while Figure 5.6 depicts the residuals of the median regression.

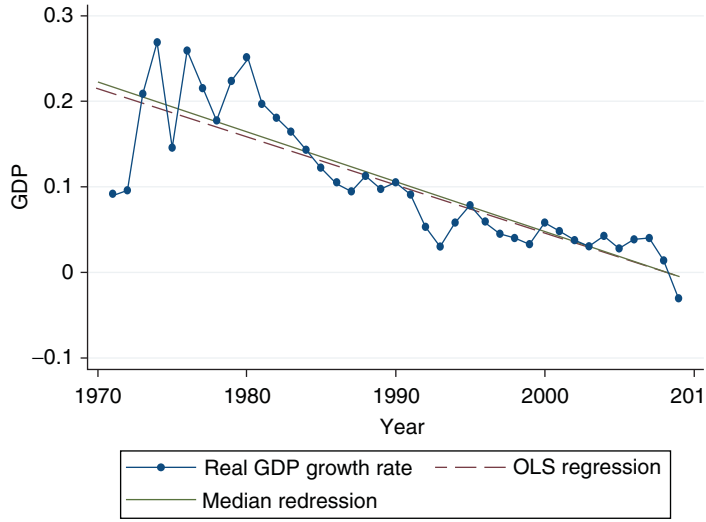

Figure 5.5 OLS and median regression for the growth rate of Italian GDP*, annual data from 1971 to 2009 in a sample of size* n = *39. In the 1970s the data are more dispersed around the two fitted lines than in the other decades of the sample.*

[1] When the term $\beta(\theta_i) - \beta(\theta_j)$ is a vector involving more than one comparison at a time, $\mathbf{\Delta}$ will have 1 on the main diagonal and -1 on the diagonal to its right-hand side, while $\mathbf{\Psi}$ embeds a $p-1$ identity matrix in order to exclude the intercept.

[2] Source: Istat, Conti economici nazionali.

Table 5.2 Median regression of the Italian GDP growth rate.

	Coefficient	Standard error	Student-t	95% idence interval	
Year	−0.0057947	0.0005357	−10.82	−0.00688	−0.0047093
Constant	11.63824	1.065982	10.92	9.478356	13.79813

Sample size $n = 39$, annual data from 1971 to 2009.

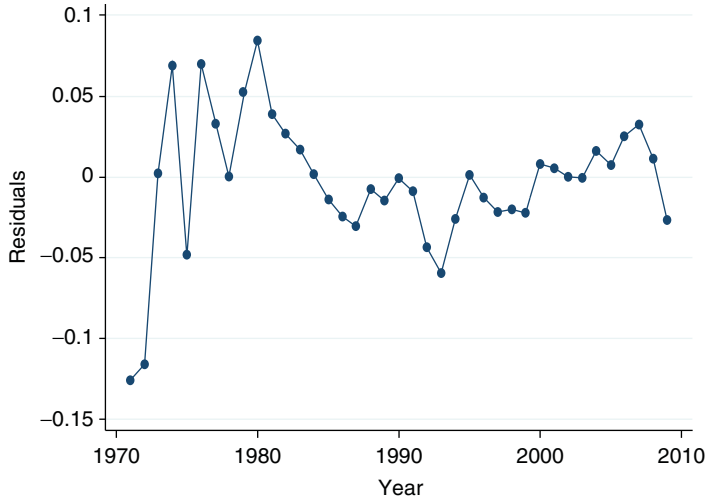

Figure 5.6 Residuals of the median regression for the GDP *rate of growth equation, sample size* n = *39. The residuals in the first decade are more dispersed than in the rest of the sample.*

The graph in Figure 5.6 shows a first period characterized by higher volatility, followed by a more stable period. While the example in Figure 5.4 presented a case of increasing heteroskedasticity over time, this is instead a model where heteroskedasticity decreases over time: the variance in the first part of the sample is higher than at the end of the sample.

A simple OLS test for heteroskedasticity is an F test: the sample is split in two subgroups, leaving out a few central observations to ensure independence of the two subsets, and the equation is independently estimated in the two subsamples. The test function is given by the ratio of the sum of square residuals of the two regressions.[3] The OLS estimates in the two subgroups, from 1971 to 1982 and from 1990 to 2009, leaving out a middle group of data to grant independence between the estimates in the two different subsets, yield, respectively, the estimated variances $\widehat{\sigma_1} = 0.003028$ and $\widehat{\sigma_2} = 0.000365$. The test is $F = 4.609$, to be compared with the critical values of

[3] This is the well known Goldfeld and Quandt (1965) test, applicable when one explanatory variable can be considered as the main source of heteroskedasticity.

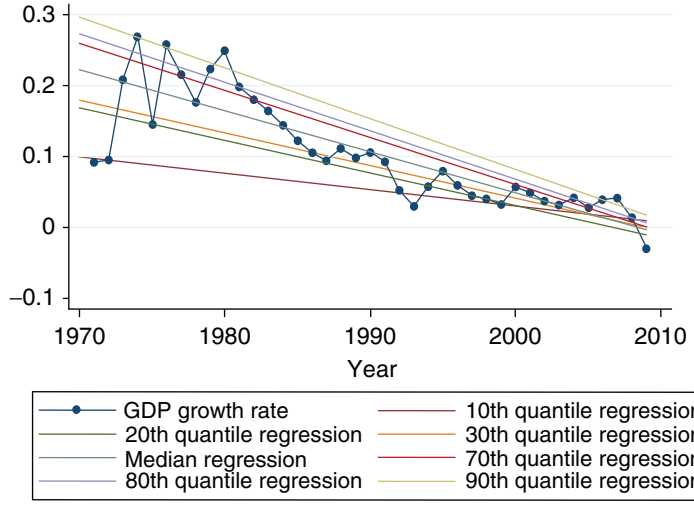

Figure 5.7 Estimated QRs for the Italian GDP *growth rate equation, sample size* n = *39. The variance at the beginning of the sample is sizeably larger than the one at the end of the sample.*

$F_{(n_1-p),(n_2-p)} = F_{10,18} = 2.41$ for $\alpha = 5\%$ and $F_{10,18} = 3.51$ for $\alpha = 1\%$. The null of equal variances in the two subsamples, as estimated by OLS, is rejected.

A first quick check for the presence of heteroskedasticity in the QR setting can proceed analogously. The estimate of the sparsity function at the median in the first subset yields the value $\hat{\omega}^2(0.50)_{1971-1982} = (4 * 4.90619442)^{-1} = 0.050956$, while in the second subset the estimated nuisance parameter at the median is $\hat{\omega}^2(0.50)_{1990-2009} = (4 * 14.9432292)^{-1} = 0.016729$. These results do indeed mirror a larger nuisance in the first subset, but a formal test has to be implemented in order to asses the presence of heteroskedasticity in the data.[4]

As mentioned, heteroskedasticity has an impact on the estimated slope coefficients at the different QRs. Figure 5.7 and Table 5.3 consider the GDP growth rate as a function of an intercept plus a time trend at the 10th, 20th, 30th, 50th, 70th, 80th and 90th quantile. In the table, the estimated slopes are inversely related to the quantiles. At the lower quantile the estimated slope is not significantly different from zero: the dispersion at the first quantile is large enough to invalidate the presence of a time trend in the dependent variable.

Next step is to verify if the differences between the slopes estimated at the various quantiles are statistically significant or if they are irrelevant and any discrepancy is simply due to a sample effect. Looking at Figure 5.8, the farthest estimates of the

[4] In principle, analogously to the OLS analysis, it could be possible to compare the estimated objective functions independently computed in the two subsamples: the sum of the absolute value of residuals is lower in the case of low variability since the data are closer to the fitted line. In the GDP example the objective function in the first subset is equal to 0.54 while in the second subset is 0.28.

Table 5.3 QR estimates of real GDP growth rate model.

θ-quantile	0.10	0.20	0.30	0.50	0.70	0.80	0.90
Year	−0.0023	−0.0045	−0.0046	−0.0057	−0.0066	−0.0068	−0.0071
Student-t	(1.86)	(3.50)	(4.54)	(5.91)	(6.94)	(9.76)	(12.86)
Constant	4.66	9.17	9.25	11.63	13.36	13.68	14.42
Student-t	(1.88)	(3.52)	(4.57)	(5.94)	(6.98)	(9.82)	(12.94)

Absolute value of Student-t statistics in parentheses, sample size $n = 39$.

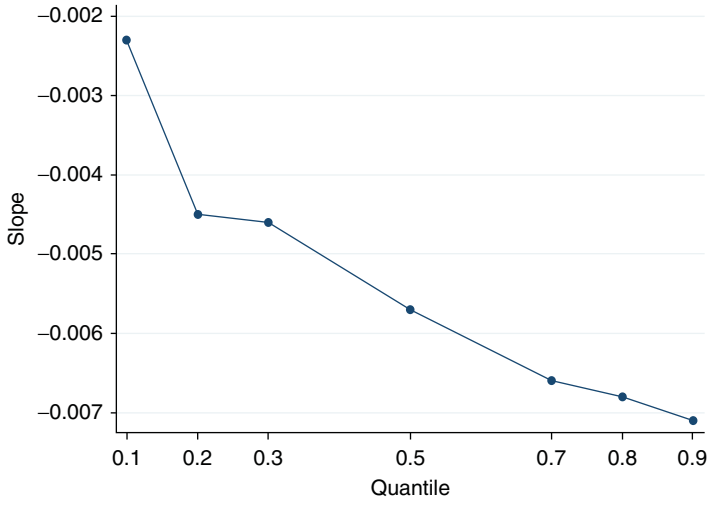

Figure 5.8 Pattern of the estimated slope at the different QRs in the GDP *growth rate equation, sample size* n = *39. The slope is inversely related to the selected QR.*

slope coefficient β are at the 10th and 90th quantiles. Thus consider estimating only these two quantiles in the entire sample and evaluating the T–test of Equation (5.11). The difference in the slopes is $(\widehat{\beta}_1(0.10) - \widehat{\beta}_1(0.90)) = -0.0023 + 0.0071 = 0.0048$. The matrix $\left[\Delta\Omega\Delta^\top \otimes \Psi\mathbf{D}^{-1}\Psi^\top\right]^{-1}$ when comparing only two quantiles coincides with $\left[var(\widehat{\beta}_1(0.10)) + var(\widehat{\beta}_1(0.90)) - 2cov(\widehat{\beta}_1(0.10), \widehat{\beta}_1(0.90))\right]^{-1} = (0.0000015 + 0.00000031 - 2 * 0.00000020)^{-1} = (0.0000014)^{-1}$, to be multiplied by $(\widehat{\beta}_1(0.10) - \widehat{\beta}_1(0.90))^2 = 0.0048^2$. The estimated value of the test is $T = 16.45$, to be compared with the critical values of a χ^2 with 1 degree of freedom, given by 3.841 for $\alpha = 5\%$ and 6.635 for $\alpha = 1\%$. The null of equal slope at the 10th and 90th QRs is rejected.

To implement the pairwise comparisons of the slope coefficient estimated at all the quantiles considered in Table 5.3, the vector of differences between the slope coefficient estimated at the nearby quantiles, $\widehat{\beta}_1(\theta_i) - \widehat{\beta}_1(\theta_j)$, the variance covariance

matrix of the slope coefficient estimated at the different quantiles, $var(\widehat{\beta}_1(\theta_i) - \widehat{\beta}_1(\theta_j))$, and the selection matrix Δ are given by:

$$\widehat{\beta}_1(\theta_i) - \widehat{\beta}_1(\theta_j) = \begin{bmatrix} \widehat{\beta}_1(0.20) - \widehat{\beta}_1(0.10) \\ \widehat{\beta}_1(0.30) - \widehat{\beta}_1(0.20) \\ \widehat{\beta}_1(0.50) - \widehat{\beta}_1(0.30) \\ \widehat{\beta}_1(0.70) - \widehat{\beta}_1(0.50) \\ \widehat{\beta}_1(0.80) - \widehat{\beta}_1(0.70) \\ \widehat{\beta}_1(0.90) - \widehat{\beta}_1(0.80) \end{bmatrix} = \begin{bmatrix} 0.0022 \\ 0.0001 \\ 0.0011 \\ 0.0009 \\ 0.0002 \\ 0.0003 \end{bmatrix}$$

	0.10	0.20	0.30	0.50	0.70	0.80	0.90
0.10	$1.912e^{-06}$						
0.20	$1.181e^{-06}$	$2.274e^{-06}$					
0.30	$9.466e^{-07}$	$1.605e^{-06}$	$1.661e^{-06}$				
0.50	$5.884e^{-07}$	$7.666e^{-07}$	$8.756e^{-07}$	$7.231e^{-07}$			
0.70	$3.549e^{-07}$	$4.078e^{-07}$	$4.725e^{-07}$	$3.566e^{-07}$	$4.798e^{-07}$		
0.80	$1.413e^{-07}$	$3.829e^{-07}$	$4.801e^{-07}$	$2.711e^{-07}$	$3.269e^{-07}$	$4.137e^{-07}$	
0.90	$1.544e^{-07}$	$3.703e^{-07}$	$3.238e^{-07}$	$1.807e^{-07}$	$2.377e^{-07}$	$3.149e^{-07}$	$3.692e^{-07}$

$$\mathbf{\Delta} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

The above matrices lead to computing the test function for the pairwise comparison of the slope estimated at all the selected QRs. The T-test assumes the value $T = 14.11$, to be compared with the critical values of a χ^2_6. With 6 degrees of freedom and $\alpha = 5\%$ the critical value is 12.59, while for $\alpha = 1\%$ the critical value is 16.81. The decision is ambiguous, the null is rejected at the 5% level but is not rejected when choosing $\alpha = 1\%$. While the comparison of the 10th and 90th quantiles provides a clear answer, the comparison of adjacent quantiles yields a more ambiguous outcome.

5.2.2 Quick ways to test equality of coefficients across quantiles in Stata

Stata provides a couple of simple instructions to verify the null of equality of the coefficients across quantiles. Instead of individually estimating the model at the given quantiles, the instruction "sqreg *GDP* year, q(0.1 0.2 0.3 0.5 0.7 0.8 0.9)" provides the simultaneous estimate of the equation at more than one quantile and, above all, provides the estimated variance-covariance matrix across quantiles.

This allows to quickly compute tests on the coefficients at different quantiles. For instance, to verify the null H_0: $\beta_1(0.20) - \beta_1(0.90) = 0$, the test is implemented by writing "test [q20]year=[q90]year" immediately after the "sqreg" instruction. The estimated test function is $T = 3.52$ to be compared with the critical value of $F_{(1,36)} = 4.11$ at $\alpha = 5\%$ and of $F_{(1,36)} = 7.39$ at $\alpha = 1\%$, so that the null cannot be rejected.[5] To test the null H_0: $\beta_1(0.10) - \beta_1(0.90) = 0$, the result is $T = 10.24$ and the null is rejected, thus confirming the previous results.

A quick way to implement the pairwise comparisons of slopes computed at nearby quantiles, as shown at the end of Section 5.2.1, is given by repeatedly implementing the "test" command (see Appendix 5.A.3). This provides the result $T = 2.97$ to be compared with the critical values of $F_{(6,36)} = 2.36$ at $\alpha = 5\%$ and $F_{(6,36)} = 3.35$ at $\alpha = 1\%$. Once again the null of equality of coefficients is rejected at the 5% but not at the 1% level when comparing the slope computed at all the adjacent QRs.

To test the single hypothesis H_0: $\beta_1(0.90) - \beta_1(0.10) = 0$, the instruction "lincom [q90]year-[q10]year" after the regression command "sqreg" can be implemented as well. It reports t-statistics and confidence intervals of the linear combination of the coefficients of the two QRs (see Appendix 5.A.3). The estimated difference between the slope computed at these two quantiles is -0.0048 with a Student-t value of 3.20 and the null of equality is rejected, while the test H_0: $\beta_1(0.90) - \beta_1(0.20) = 0$ yields an estimated difference of -0.0026 with a student-t value of 1.88 and the null is not rejected.

A different instruction is the "iqreg" command, which provides the inter-quantile difference, repeatedly implemented for this same purpose in Chapter 3. In the GDP model the instruction "iqreg *GDP* year, q(0.20 0.90)" considers the difference between the 90th and the 20th quantiles:

$$[GDP(0.90) - GDP(0.20)] = [\beta_0(0.90) - \beta_0(0.20)] + [\beta_1(0.90) - \beta_1(0.20)]\, year. \tag{5.12}$$

The estimated difference in the slope is $\widehat{\beta}_1(0.90) - \widehat{\beta}_1(0.20) = -0.00026$, with standard error $se\left[\widehat{\beta}_1(0.90) - \widehat{\beta}_1(0.20)\right] = 0.0014$ and Student-t value of -0.185, which does not reject the hypothesis of equality of the two slopes. The distance between the 10th and 90th quantile as measured by the interquantile difference is $\widehat{\beta}_1(0.90) - \widehat{\beta}_1(0.10) = -0.0048$ with a Student t-value of -3.92, once again rejecting the null of equality.

The advantage of "lincom"and "iqreg"is not in the estimates, but in the Student-t tests, which immediately provide results on the equality of coefficients across quantiles. The pairwise comparisons of the slopes of close by quantiles for Equation (5.12), instead, do not reject the null of equality. Indeed, the estimated differences between adjacent quantiles, as computed by "iqreg", are not significantly different from zero:

[5] The "sqreg" instruction implements bootstrap to compute the variance-covariance matrix. This is why the results may be numerically different.

$$\begin{aligned}
\widehat{\beta}_1(0.20) - \widehat{\beta}_1(0.10) &= -0.0022 \qquad t = -1.21 \\
\widehat{\beta}_1(0.30) - \widehat{\beta}_1(0.20) &= -0.000033 \qquad t = -0.04 \\
\widehat{\beta}_1(0.50) - \widehat{\beta}_1(0.30) &= -0.0011 \qquad t = -1.08 \\
\widehat{\beta}_1(0.70) - \widehat{\beta}_1(0.50) &= -0.00085 \qquad t = -1.37 \\
\widehat{\beta}_1(0.80) - \widehat{\beta}_1(0.70) &= -0.00015 \qquad t = -0.28 \\
\widehat{\beta}_1(0.90) - \widehat{\beta}_1(0.80) &= -0.00036 \qquad t = -0.62.
\end{aligned}$$

Summarizing, although numerically different, all the tests above implemented yield the same results: the 10th and the 90th quantile are significantly different from one another while the 20th and the 90th are not. The pairwise comparisons across all quantiles yield more ambiguous results, rejecting the equality at the 5% but failing to reject at the 1% significance level.

The tests discussed in this section are Student-t or F-tests. The Student-t test is the usual result provided in the output of a regression model, in this case measuring the inter-quantile difference, and it is computed by the ratio of the estimated coefficient and its own standard error. The F-test, instead, is the result of an approximation. As discussed in Section 3.3, Wald, likelihood ratio and Lagrange multiplier tests verify the validity of constraints involving the exclusion of more than one coefficient at a time. They can also be implemented to test constraints across different QRs. The three tests are asymptotically equivalent and are distributed as χ^2 with degrees of freedom r which are equal to the number of constraints under test. These tests can be approximated by rF plus a remainder. Stata takes advantage of this approximation to implement the "test" instruction.

5.2.3 The wage equation revisited

In light of the analysis developed in Section 5.2, it is now possible to reconsider the Italian wage equation analyzed in Chapter 3.2. Instead of a subsample of size $n = 49$ now all the data of year 2006 in the Survey of Household Income and Wealth is analyzed, and this involves a sample of size $n = 5225$. To make results comparable with those of Table 3.3 and Table 3.4, the simple linear regression model is computed. The dependent variable is log of *wage* and the explanatory variable is *education* in its four degrees: *elementary*, *junior high*, *high school* and *university*.

Figure 5.9 presents the density of *lwage* [Figure 5.9(a)] and its quantile function, $F^{-1}(\theta)$, compared with the uniform and normal quantile functions [Figure 5.9(b) and (c), respectively]. The density is positively skewed, and the quantile function shows a sharp increase after the third quartile. Table 5.4 provides the summary statistics of earnings and, compared with Table 3.2, *lwage* presents a decreased variability and a sizeable increase in skewness. Table 5.5 presents the QR estimates, with a slope ranging from 0.025 to 0.064. Thus the educational premium at the higher wages/quantiles is more than twice than returns at the lower wages/quantile. However in Table 3.3 the estimated slope assumes values ranging from 0.068 to 0.091, thus in the larger sample education premia are smaller and more dispersed.

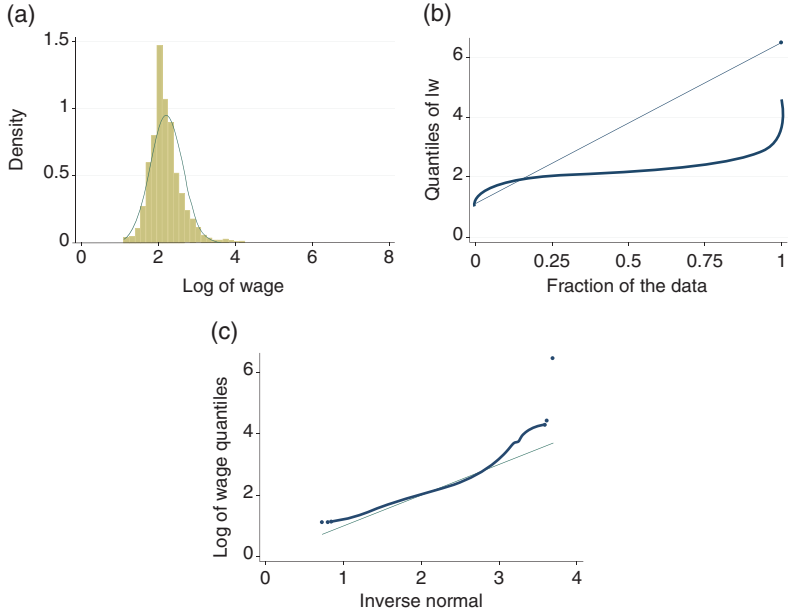

Figure 5.9 Density function (a) of the dependent variable, log of wage: the solid line is the normal density. (b) and (c) depict the Quantile plots of the dependent variable, log of wage*: the straight line in (b) represents the equal distribution case, the uniform quantile function, while it represents the normal quantile function in (c).*

Table 5.4 Summary statistics for the log of *wage* dependent variable

	Mean	Median	Standard deviation	25th percentile	75th percentile	Skewness
lwage	2.207	2.151	0.148	1.950	2.392	1.256

Table 5.5 QR estimates for the log of *wage* equation.

θ-quantile	0.10	0.25	0.50	0.75	0.90
education	0.0259	0.0287	0.0368	0.0512	0.0637
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
intercept	1.495	1.643	1.745	1.799	1.888
	(0.025)	(0.029)	(0.023)	(0.019)	(0.033)

Standard errors in parentheses, sample size $n = 5225$.

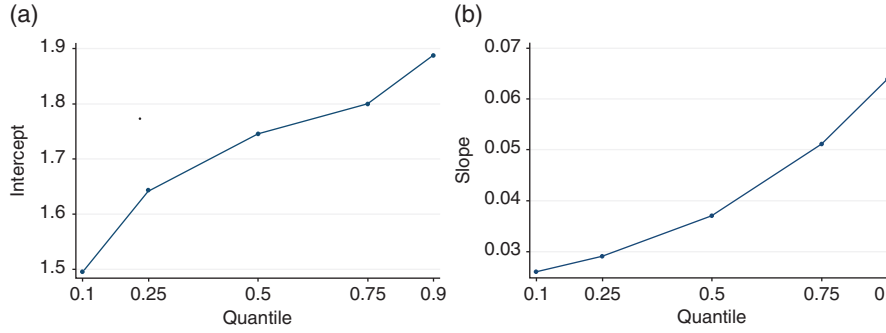

Figure 5.10 Values of the regression coefficients [the intercept in (a) and the slope in (b)] computed at different QRs. Wage equation, year 2006, sample size n = *5225.*

The pattern of the estimated intercept and slope at the various quantiles are depicted in Figure 5.10.

Table 5.6 presents the interquantile differences. This is the most relevant discrepancy between the results provided by the small and large samples. In the large sample the interquantile differences are all statistically different from zero, with the sole exception of the comparison between the slope at the 10th and 25th quantile. Conversely in the small sample all the interquantile differences are statistically irrelevant. Thus the wage equation estimated in the large sample is a location and scale shift model. Figure 5.11 displays the estimated QRs where the fitted lines are not parallel, in contrast to what occurred in Figure 3.21 for the small sample. In particular, the differences in slope are $\widehat{\beta}_1(0.5) - \widehat{\beta}_1(0.25) = 0.008$ and $\widehat{\beta}_1(0.75) - \widehat{\beta}_1(0.5) = 0.014$, showing that the increase in returns to education at the third quartile with respect to the median almost doubles the increase of returns at the median with respect to the first quartile. The discrepancy is further enhanced when comparing $\widehat{\beta}_1(0.50) - \widehat{\beta}_1(0.10) = 0.011$ and $\widehat{\beta}_1(0.90) - \widehat{\beta}_1(0.50) = 0.027$. This leads to the conclusion that higher degrees of education grant increasing returns.

Finally, testing the equality of the slope estimated in the first two QRs yields a value of $T = 2.73$ and the null $H_0: \beta_1(0.10) = \beta_1(0.25)$ is not rejected when compared with the critical value of $F_{(1,\,5223)} = 3.84$. The test of equality of the slope across all quantiles, instead, yields the value $T = 144.12$ and the null $H_0: \beta_1(0.10) = \beta_1(0.25) = \beta_1(0.50) = \beta_1(0.75) = \beta_1(0.90)$ is rejected when compared with the critical value of $F_{(4,5223)} = 2.37$.

However in Section 5.1.3 the wage equation has two explanatory variables, *education* and *age*, and it is of course possible to reintroduce the *age* variable into the model. The estimates for the large sample are reported in Table 5.7, where the coefficients are all statistically relevant. In Table 5.8, which collects the interquantile differences, the discrepancies in the intercept are mostly irrelevant while the differences in both slope coefficients are significantly different from zero, once again confirming the presence of heteroskedasticity. A graphical analysis can be

Table 5.6 Interquantile differences in the estimated coefficients of the wage equation, year 2006.

	0.25−0.10	0.50−0.25	0.75−0.50	0.90−0.75	0.90−0.10	0.50−0.10	0.90−0.50
education	0.0028	0.0080	0.0144	0.0125	0.0378	0.0108	0.0269
	(0.002)*	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
intercept	0.1480	0.1019	0.0544	0.0887	0.3931	0.2499	0.1432
	(0.021)	(0.025)	(0.024)	(0.027)	(0.032)	(0.028)	(0.035)

Standard errors in parentheses, sample size $n = 5225$. The asterisk indicates the estimates characterized by low Student-t values.

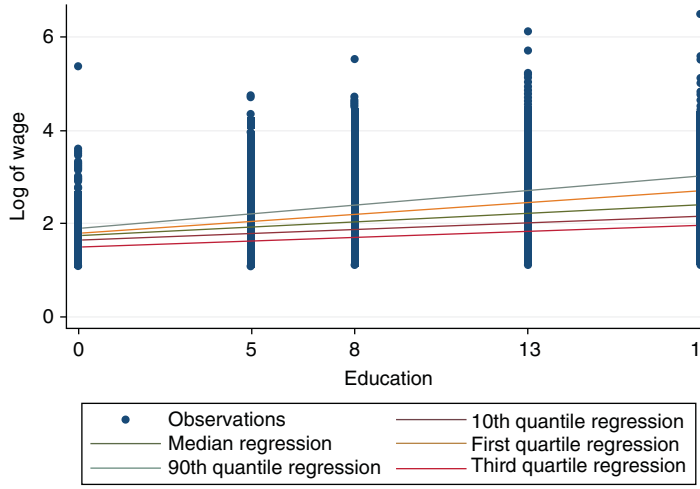

Figure 5.11 QR estimates of the wage equation in Italy, year 2006. In the large sample the estimated QRs are no longer parallel lines.

Table 5.7 QR estimates of the wage equation, year 2006.

θ-quantile	0.10	0.25	0.50	0.75	0.90
education	0.0344	0.0369	0.0435	0.0532	0.0611
	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
age	0.0118	0.0106	0.0127	0.0155	0.0174
	(0.0006)	(0.0005)	(0.0004)	(0.0007)	(0.001)
intercept	0.9122	1.111	1.133	1.110	1.154
	(0.042)	(0.035)	(0.026)	(0.019)	(0.057)

Standard errors in parentheses, sample size $n = 5225$.

Table 5.8 Interquantile differences in the estimated coefficients of the wage equation, year 2006.

	0.25−0.10	0.50−0.25	0.75−0.50	0.90−0.75
education	0.0025	0.0065	0.0096	0.0079
	(0.002)∗	(0.001)	(0.001)	(0.002)
age	−0.0012	0.0020	0.0028	0.0019
	(0.0006)	(0.0005)	(0.0006)	(0.0008)
intercept	0.1988	0.0227	−0.0231	0.0437
	(0.038)	(0.040)∗	(0.035)∗	(0.046)∗

Standard errors in parentheses, sample size $n = 5225$. The asterisks indicate the estimates characterized by low Student-t values.

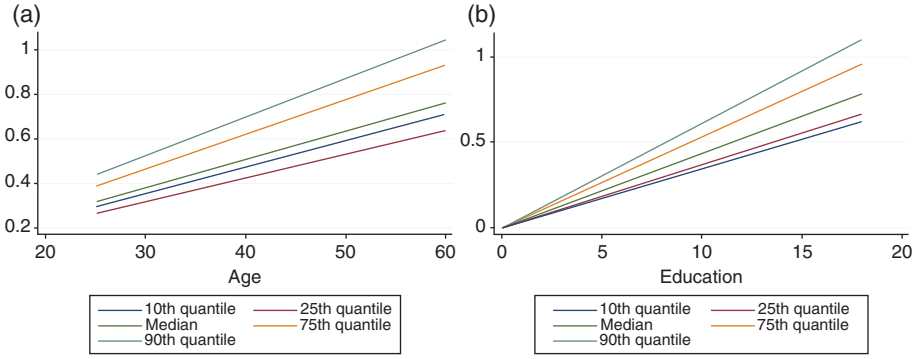

Figure 5.12 The lines depict the contribution of the explanatory variable age *(a) and* education *(b) to the dependent variable, log of* wage*, at the selected quantiles in the large sample of size* n = *5225. These lines are not parallel.*

implemented by considering the role of each independent variable in explaining the dependent variable log of *wage*. The individual role can be measured by multiplying the coefficient estimated at a given quantile by its explanatory variable, for instance 0.0344**education* computes the role of education in explaining log of *wage* at the 10th quantile, while 0.0118**age* yields the role of age, as a proxy of experience, to explain log of *wage* at the 10th quantile. Figure 5.12 reports the individual share of each independent variable at the different quantiles and, as expected, these lines are not parallel and rather clearly describe a scale shift model.

5.3 The dependent data model

In this section the assumption of independent observations is abandoned. This is typically a time series problem: the errors are correlated and the current value of the dependent variable is influenced by its own past values.

The link existing between recent and past observations can be ascribed to different causes. Besides a genuine link between the dependent variable and its previous values, serial correlation can be induced by an incorrect specification of the estimated regression. This is particularly true if the estimated equation excessively simplifies the dynamics of the model by omitting lagged dependent and/or lagged explanatory variables, as will be seen shortly.

Consider the equation $y_i = \beta_0 + \beta_1 x_i + e_i$, to be estimated at the quantile θ in a sample of size n. In the simplest specification the errors are identically distributed but not independent, and are defined by a first order autoregressive process $e_i = a\ e_{i-1} + \mathrm{a}_i, e_i \sim AR(1)$, under the assumptions: $|a| < 1$, e_i having common positive density $f, f(F^{-1}(\theta)) > 0$, and a_i being i.i.d.

The condition $|a|<1$ implies that past errors do influence y_i, but that their impact decreases over time. Indeed, substituting e_i with $a\ e_{i-1} + \mathrm{a}_i$, which is its $AR(1)$ definition, the model becomes:

$$y_i = \beta_0 + \beta_1 x_i + a e_{i-1} + \mathrm{a}_i, \tag{5.13}$$

in turn, by replacing in (5.13) e_{i-1} with its $AR(1)$ definition, $e_{i-1} = a\, e_{i-2} + \mathrm{a}_{i-1}$ one has:

$$y_i = \beta_0 + \beta_1 x_i + a^2 e_{i-2} + a\mathrm{a}_{i-1} + \mathrm{a}_i = \beta_0 + \beta_1 x_i + a^2 e_{i-2} + \lambda_i,$$

where $\lambda_i = a\mathrm{a}_{i-1} + \mathrm{a}_i$. By repeatedly replacing the lagged error term with its $AR(1)$ definition, that is going back h periods, it is:

$$y_i = \beta_0 + \beta_1 x_i + a^h e_{i-h} + a^{h-1}\mathrm{a}_{i-h+1} + \ldots\ldots + a^2\mathrm{a}_{i-2} + a\mathrm{a}_{i-1} + \mathrm{a}_i =$$
$$= \beta_0 + \beta_1 x_i + a^h e_{i-h} + \lambda_i,$$

where $\lambda_i = a^{h-1}\mathrm{a}_{i-h+1} + \ldots\ldots + a^2\mathrm{a}_{i-2} + a\mathrm{a}_{i-1} + \mathrm{a}_i$ gathers all the actual and past values of a_i. In the above equation the far away error $a^h e_{i-h}$ has very little influence on current y_i as long as $|a|<1$. This condition is needed to ensure the possibility to estimate the model. The term $\lambda_i = a^{h-1}\mathrm{a}_{i-h+1} + \ldots\ldots + a^2\mathrm{a}_{i-2} + a\mathrm{a}_{i-1} + \mathrm{a}_i$, being the sum of i.i.d. errors, is itself i.i.d.

Weiss (1990) first considers the median regression to estimate and test a serial correlation model. If the model is estimated ignoring the presence of serial correlation, the estimates are unbiased but inference is incorrect. The estimated nuisance parameter underestimates or overestimates the true value depending on the sign of the neglected error correlation. The QR objective function $V(\theta)$ and the gradient $V'(\theta)$, in turn for the intercept and the slope, are the following:

$$V(\theta) = \sum_{i=1,..n} \rho(y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1,..n} \rho(e_i);$$
$$V'(\theta) = \sum_{i=1,..n} \psi(y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1,..n} \psi(e_i) = 0$$
$$\sum_{i=1,..n} \psi(y_i - \beta_0 - \beta_1 x_i) x_i = \sum_{i=1,..n} \psi(e_i) x_i = 0.$$

At the median $\psi(e_i) = sgn(e_i)$ and the quadratic form of the gradient converges to the positive definite matrix:

$$\mathbf{A} = lim_{n\to\infty} 1/n \sum_i \sum_j \psi(e_i)\psi(e_j)\mathbf{x}_i\mathbf{x}_j^\top,$$

where $\mathbf{x}_i = \begin{bmatrix} 1 \\ \mathrm{x}_i \end{bmatrix}$ is the $(p, 1)$ vector collecting the i-th observation of all the explanatory variables in the equation. Further assuming that the independent variable has quadratic form converging to the positive definite matrix $\mathbf{D} = lim_{n\to\infty} 1/n \sum_i \mathbf{x_i}\ \mathbf{x_i}^\top$, the QR estimator minimizing $V(\theta)$ is asymptotically distributed as:

$$\sqrt{n}[\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)] \to N(0, \omega^2(\theta)\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}). \tag{5.14}$$

Ignoring serial correlation means to incorrectly estimate the term $\omega^2(\theta)\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ by the term $\omega^2(\theta)\mathbf{D}^{-1}$ as in Equation (5.1).

To improve efficiency, the equation $y_i = \beta_0 + \beta_1 x_i + a\ e_{i-1} + \mathrm{a}_i$ is transformed by replacing the lagged error term e_{i-1} with its definition $y_{i-1} - \beta_0 - \beta_1 x_{i-1}$. The model becomes:

$$y_i = \beta_0 + \beta_1 x_i + a(y_{i-1} - \beta_0 - \beta_1 x_{i-1}) + \mathrm{a}_i$$

$$y_i = (1 - a)\beta_0 + a y_{i-1} + \beta_1 (x_i - a x_{i-1}) + \mathrm{a}_i$$

$$y_i = b_0 + a y_{i-1} + \beta_1 x_i - b_1 x_{i-1} + \mathrm{a}_i \tag{5.15}$$

$$y_i - a y_{i-1} = b_0 + \beta_1 (x_i - a x_{i-1}) + \mathrm{a}_i$$

$$y_i^* = \beta_0 + \beta_1 x_i^* + \mathrm{a}_i, \tag{5.16}$$

where in Equation (5.15) $b_1 = a\beta_1$ and $b_0 = (1 - a)\beta_0$, while in Equation (5.16) y_i^* and x_i^* are defined as $y_i^* = y_i - a y_{i-1}$ and $x_i^* = (x_i - a x_{i-1})$.

Equation (5.15) and Equation (5.16) allow to estimate the model with i.i.d. errors, since their variables have been transformed, or additional lagged variables have been introduced, in order to have a_i as the error term of the equation, which is i.i.d. by assumption.

With respect to the initial equation which ignores the presence of serial correlation, $y_i = \beta_0 + \beta_1 x_i + e_i$, the model in (5.15) involves the introduction of lagged dependent and independent variables. This explains the initial statement at the beginning of this section, that serial correlation can be due to an incorrect definition of the dynamic of the model, that is to the omission of lagged variables. The vector of regression coefficients now includes the correlation coefficient a individually, in the coefficient of the lagged dependent variable $a y_{i-1}$, and in a multiplicative way in both the constant term, $b_0 = (1 - a)\beta_0$, and in the coefficient of the lagged independent variable, $b_1 x_{i-1} = a\beta_1 x_{i-1}$. The vector of the regression coefficient is now $\boldsymbol{\beta}^\top(\theta) = [b_0(\theta)\ a(\theta)\ \beta_1(\theta)\ b_1(\theta)]$.[6] Equation (5.15) provides a quick way to verify the presence of serially correlated errors. The t-test on the estimated coefficient of the lagged dependent variable, $a(\theta)$, allows to accept or reject the null $H_0 : a = 0$, which implies absence of correlation. Thus the test for serial correlation is a t-test on the lagged dependent variable in Equation (5.15).

Model (5.16), instead, is a two step estimator. In the first step the correlation coefficient a is computed by $a = cov(e_i e_{i-1})/var(e_i)$, where e_i, are replaced by the residuals of the initial equation, $y_i = \beta_0 + \beta_1 x_i + e_i$. In the second step the original variables are transformed into $y_i^* = y_i - a\ y_{i-1}$ and $x_i^* = (x_i - a x_{i-1})$ in

[6] In the case of $e_i = AR(2) = a_1 e_{i-1} + a_2 e_{i-2} + \mathrm{a}_i$, two lags of both dependent and independent variables will be included in the final equation. In the case of higher order serial correlation, $e_i = AR(q) = a_1 e_{i-1} + a_2 e_{i-2} + \ldots + a_q e_{i-q} + \mathrm{a}_i$, q lags must be considered.

order to purge correlation. The vector of the coefficient in the model of Equation (5.16) is $\boldsymbol{\beta}^{\top}(\theta) = [\beta_0(\theta)\ \beta_1(\theta)]$.[7]

After purging serial correlation, the asymptotic distribution of the QR estimator is given by $\sqrt{n}\left[\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)\right] \to N(0, \omega^2(\theta)\mathbf{A}^{-1})$. The efficiency gain implied by model (5.16) can be found in Weiss (1990), who shows that $\omega^2(\theta)\mathbf{A}^{-1} < \omega^2(\theta)\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$.

5.3.1 Example with dependent data

This section analyzes the dataset on the spot exchange rate between the German mark and the US dollar at delivery. This dataset considers observations before the introduction of the euro, and relates the spot exchange rate to the forward exchange rate (Hayashi 2000). The equation provides an estimate of market efficiency. The idea behind this model is to quantify to what extent the changes in the 30-day forward realizations of the exchange rate, that is the spot rates, take into account the forecasts of that same rate, that is the forward rates. Under the hypotheses of rationality and risk neutrality, the economic theory states that the forward rate is the optimal forecast of future spot rates. The dataset is given by weekly data from January 1975 to September 1987 with a sample of size $n = 664$. The independent variable is given by the log of a 30-day forward mark–dollar exchange rate, $lf30 = log(f30)$. The dependent variable of the regression is the log of the spot rate of a 30-day forward contract on delivery, $ls30 = log(s30)$. The equation considers whether the changes in the spot rate are a function of the forward rates, in order to verify the hypotheses of rationality and risk neutrality. These hypotheses imply that the forward rate is the optimal forecast of future spot rates.

The estimated median regression is:

$$log(s30) = \underset{(0.007)}{0.0059} + \underset{(0.009)}{0.995}\, log(f30), \tag{5.17}$$

with standard errors in parentheses.

The estimated density at the median of the residual is given by $\widehat{f(F(0.50)^{-1})} = 13.76$ and the sparsity function is $\widehat{s(0.50)} = \left[\widehat{f(F(0.50)^{-1})}\right]^{-1} = 0.0726$. The constant scale parameter of the model is $\widehat{\omega}^2(0.50) = \theta(1-\theta)/\left[\widehat{f(F(0.50)^{-1})}\right]^2 = 0.00132$. The covariance matrix $\omega^2(\theta)\mathbf{D}^{-1}$ is estimated by $\begin{bmatrix} 0.00005361 & -0.0000627 \\ -0.0000627 & 0.00007614 \end{bmatrix}$. The square root of the diagonal elements of this matrix provide the standard errors of the regression coefficient: $se(\widehat{\beta}_0(0.50)) = 0.00732$ for the intercept and $se(\widehat{\beta}_1(0.50)) = 0.00872$ for the slope parameter.

[7] In the case of $e_i = AR(2)$, $y_i^* = y_i - a_1 y_{i-1} - a_2 y_{i-2}$, and $x_i^* = x_i - a_1 x_{i-1} - a_2 x_{i-2}$. In the case of $e_i = AR(q)$, q autoregression coefficients have to be estimated and q lags have to be introduced to compute the transformed variables y_i^* and x_i^*.

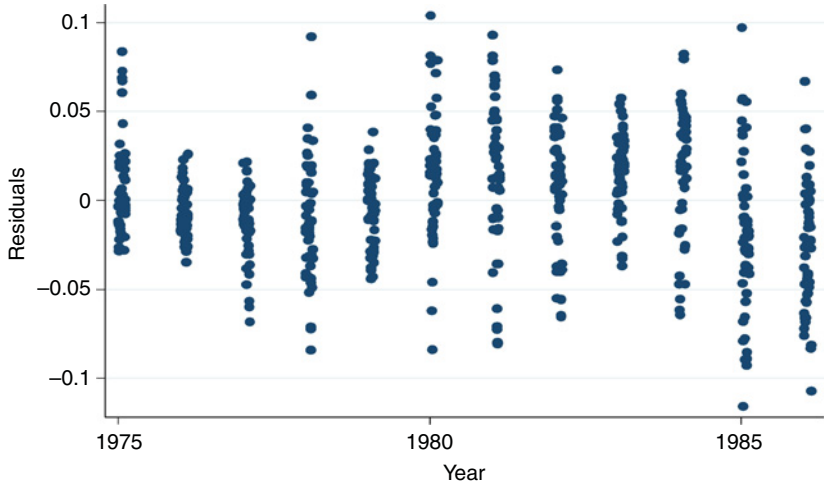

Figure 5.13 Residuals from the equation of the spot rate as a function of the forward rates, as estimated in Equation (5.17). The wavy pattern of this graph may be a signal of serial correlation in the dataset.

These results hint that the forward rate does influence the spot rate and that the rationality and risk neutrality assumptions of the economic theory are confirmed by the data.

However, with time series data it is wise to implement an additional check to control the presence of serial correlation. By looking at the residuals from Equation (5.17), reported in Figure 5.13, a systematic wavy pattern can be traced and this is a typical pattern of correlated errors. The residuals from the estimated regression in (5.17) are then used to compute the first order serial correlation coefficient, estimated by $\widehat{a} = cov(\widehat{e}_i\widehat{e}_{i-1})/var(\widehat{e}_i)$, as in the two step approach of Equation (5.16). Actually the correlation coefficient a can also be computed by implementing an OLS regression of $\widehat{e}_i$ on $\widehat{e}_{i-1}$. The latter approach has the advantage of providing standard error and t-test to judge the validity of the estimated correlation coefficient. The resulting estimate of the correlation coefficient is $\widehat{a} = 0.812$ with standard error $se(\widehat{a}) = 0.022$, which shows the presence of a not negligible serial correlation in the model. The hypothesis of rationality and risk neutrality should be further examined since the conclusion can be affected by the presence of serial correlation. The estimated serial correlation coefficient allows to transform the data into $s_i^* = log(s30_i) - a\,log(s30_{i-1})$ and $f_i^* = log(f30_i) - a\,log(f30_{i-1})$ as in (5.16). The median regression estimated with the transformed variables $s_i^* = log(s30_i) - 0.812log(s30_{i-1})$ and $f_i^* = log(f30_i) - 0.812log(f30_{i-1})$, yields the following results:

$$s_i^* = \underset{(0.003)}{0.021} + \underset{(0.024)}{0.867} f_i^*, \tag{5.18}$$

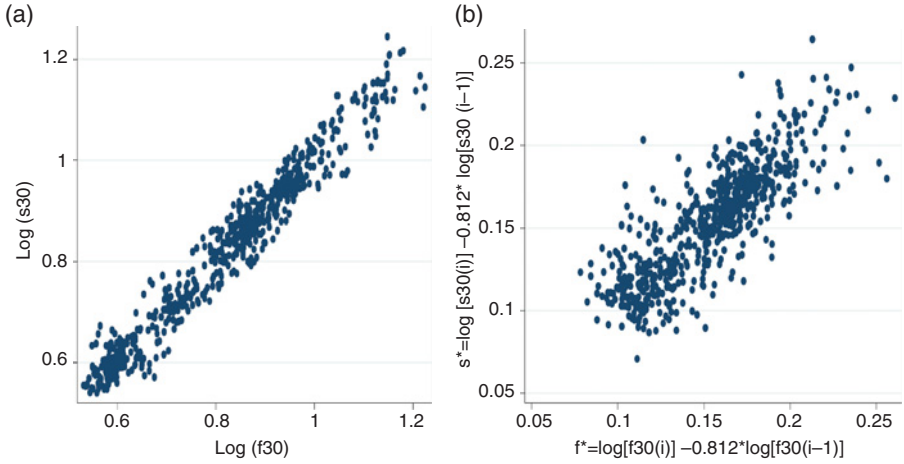

Figure 5.14 Spot rate and forward rate in a sample of size n = *664, the original dataset, in (a). In (b) are the transformed variables* f_i^* *and* s_i^*. *After purging serial correlation the data are comprised in a smaller range.*

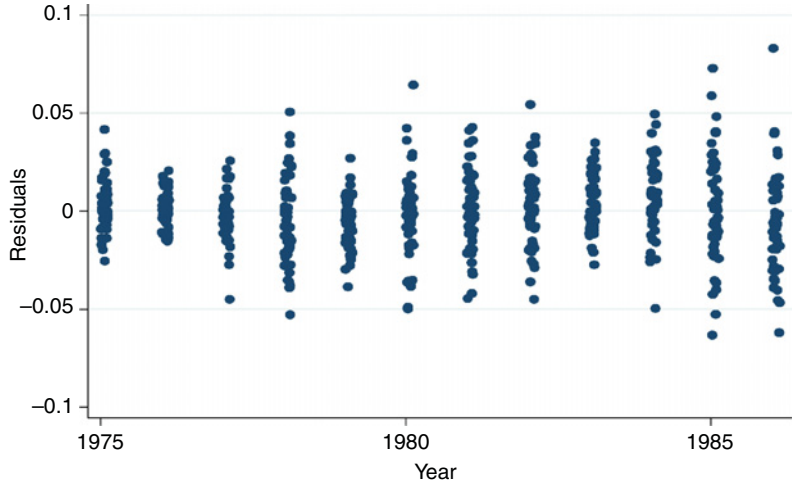

Figure 5.15 Residuals from Equation (5.18) using the transformed data s_i^* *and* f_i^*. *Purging serial correlation has curbed the wavy pattern of the residuals.*

with a slope statistically different from 0, thus confirming the rationality and risk neutrality assumption of the economic theory. Figure 5.14 reports the graph of the original data [Figure 5.14(a)] and the plot of the transformed variables [Figure 5.14(b)]. Both graphs show a clustering around an upward line. Figure 5.15 presents the residuals from Equation (5.18) computed with the transformed variables, and the graph is decidedly less wavy than the one in Figure 5.13.

5.4 Summary of key points

- Section 5.1 focuses on the estimation of the scale parameter of the equation and on the standard errors of the estimated coefficients, in order to compute confidence intervals and hypothesis testing for each estimated coefficient. This first section deals with i.i.d. errors.
- Section 5.2 considers the case of non-identical conditional distributions. Different tests to verify the hypothesis of heteroskedastic errors are discussed. An empirical example of the implementation of these tests considers annual data of the Italian GDP. Both a formal approach and some quick ways to test the i.ni.d. hypothesis are discussed.
- Section 5.3 analyzes the case of dependent data discussing two different methods to cope with dependent (ni.i.d.) errors. Both methods embed a test to verify the presence of $AR(q)$ errors. The estimation of an exchange rate equation provides the empirical example of this section.

References

Buchinsky M 1995 Estimating the asymptotic covariance matrix for quantile regression models. A Monte Carlo study. *Journal of Econometrics* **68**, 303–338.

Goldfeld S and Quandt R 1965 Some tests for heteroskedasticity. *Journal of the American Statistical Association* **60**, 539–547.

Hayashi F 2000 *Econometrics*. Princeton University Press.

Koenker R 2005 *Quantile Regression*. Cambridge University Press.

Koenker R and Basset G 1982 Robust tests for heteroskedasticity based on regression quantiles. *Econometrica* **50**, 43–61.

Koenker R and Machado 1999 Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94**, 1296–1310.

Siddiqui M 1960 Distributions of quantiles from a bivariate population. *Journal of Research of the National Bureau of Standards* **64**, 145–150.

Weiss A 1990 Least absolute error estimation in the presence of serial correlation. *Journal of Econometrics* **44**, 127–159.

Appendix 5.A Heteroskedasticity tests and weighted quantile regression, stata and R codes

5.A.1 Koenker and Basset test for heteroskedasticity comparing two quantile regressions

Stata codes

use filename.dta	input data	
sqreg y x, q(0.20 0.90)	computes both 0.20 and 0.90 quantile regressions	
matrix coeff=e(b)	saves the estimated coefficients	
scalar b20=el(coeff,1,1)		
scalar b90=el(coeff,1,3)	saves each of them individually	(****)
scalar obs=e(N)	saves the number of observations	
matrix var=e(V)	saves the covariance matrix across quantiles	
scalar var11=el(var,1,1)	individually selects each element	
scalar var12=el(var,3,1)	of the covariance matrix	(++++)
scalar var22=el(var,3,3)		
scalar dod=var11–2*var12+var22	computes $[\mathbf{\Delta\Omega\Delta}^{\top} \otimes \mathbf{\Phi D}^{-1}\mathbf{\Phi}^{\top}]$	(###)
matrix invers=inv(dod)	computes $[\mathbf{\Delta\Omega\Delta}^{\top} \otimes \mathbf{\Phi D}^{-1}\mathbf{\Phi}^{\top}]^{-1}$	
scalar diff=(b20 − b90)^2	computes the difference $(\beta(\theta_{0.20} - \beta(\theta_{0.90}))^2$	(”””””)
matrix testkb=diff*invers	computes the final test function	(. . .)
matrix list testkb	prints the result	

5.A.2 Koenker and Basset test for heteroskedasticity comparing all quantile regressions

Stata codes

sqreg growth year, q(0.10 0.20 0.30 0.50) computes all quantile regressions

At (****) instead of saving two slope coefficients at the two quantile regressions now it saves the four slopes estimated at the four quantile regressions selected

scalar b10=el(coeff,1,1) saves each estimated slope coefficient
scalar b20=el(coeff,1,3) at all the selected quantile regressions
scalar b30=el(coeff,1,5)
scalar b50=el(coeff,1,7)

At (++++) instead of saving two variances of the slope (one for each quantile regression) and one covariance across the two quantile regressions, now it saves four variances and six covariances across the selected quantile regressions

(continued)

scalar var11=el(var,1,1)	variance of the slope, 0.10 quantile
scalar var12=el(var,3,1)	covariance of the slope, 0.20 and 0.10
scalar var22=el(var,3,3)	variance of the slope, 0.20 quantile
scalar var13=el(var,5,1)	covariance of the slope, 0.30 and 0.10
scalar var23=el(var,5,3)	covariance of the slope, 0.30 and 0.20
scalar var33=el(var,5,5)	variance of the slope, 0.30
scalar var15=el(var,7,1)	covariance of the slope, 0.50 and 0.10
scalar var25=el(var,7,3)	covariance of the slope, 0.50 and 0.20
scalar var35=el(var,7,5)	covariance of the slope, 0.50 and 0.30
scalar var55=el(var,7,7)	variance of the slope, 0.50
matrix V1=(var11,var12,var13,var15)	
matrix V2=(var12,var22,var23,var25)	
matrix V3=(var13,var23,var33,var35)	
matrix VV=(V1 \ V2 \ V3 \ V5)	provides the covariance matrix
matrix delta1=(1,-1,0,0,0)	builds the matrix $\mathbf{\Delta}$
matrix delta2=(0,1,-1,0,0)	
matrix delta3=(0,0,1,-1,0)	
matrix delta4=(0,0,0,1,-1)	
matrix delta=(delta1 \ delta2 \ delta3 \ delta4)	provides $\mathbf{\Delta}$
At (###) the scalar instruction computing dod needs to be replaced with its matrix algebra analogue	
matrix dod=delta*VV*delta'	computes $\mathbf{\Delta\Omega\Delta}^{\top} \otimes \mathbf{\Phi D}^{-1}\mathbf{\Phi}^{\top}$ replacing (###)
At (”””””) the scalar difference between the slope estimated at two regression quantiles is now the vector of pairwise comparison of the slope estimated at four different quantiles	
matrix diff=(b10-b20 \ b20-b30 \ b30-b50)	computes $\beta(\theta_i)$-$\beta(\theta_j)$ replacing (”””””)
At (....) the final test function is a product among matrices	
matrix testkb=diff'*invers*diff	computes the final test replacing (....)

5.A.3 Quick tests for heteroskedasticity comparing quantile regressions

Stata codes

sqreg growth year, q(0.20 0.30 0.50 0.70 0.80 0.90)	all quantile regressions
test [q20]year=[q90]year	testing equality of two slopes, θ=0.20 and θ=0.90

sqreg growth year, q(0.10 0.20 0.30 0.50 0.70 0.80 0.90)	all quantile regressions
test [q20]year=[q10]year	testing equality of
test [q30]year=[q20]year, accum	more slopes together
test [q50]year=[q30]year, accum	
test [q70]year=[q50]year, accum	
test [q80]year=[q70]year, accum	
test [q90]year=[q70]year, accum	

Alternatively

sqreg growth year, q(0.20 0.30 0.50 0.70 0.80 0.90)	chosen quantile regressions
lincom [q90]year-[q20]year	difference between 0.90 and 0.20 quantiles

Alternatively

iqreg growth year, q(0.20 0.90)	difference between 0.90 and 0.20 quantiles
iqreg growth year, q(0.20 0.30)	difference between 0.30 and 0.20 quantiles
iqreg growth year, q(0.30 0.50)	difference between 0.50 and 0.30 quantiles
iqreg growth year, q(0.50 0.70)	difference between 0.70 and 0.50 quantiles
iqreg growth year, q(0.70 0.80)	difference between 0.80 and 0.70 quantiles
iqreg growth year, q(0.80 0.90)	difference between 0.90 and 0.80 quantiles

5.A.4 Compute the individual role of each explanatory variable to the dependent variable

Stata codes

qreg lwage education age, q(0.25)	first quartile regression
matrix firstquartile=e(b)	save estimated coefficients
scalar firstedu=el(firstquartile,1,1)	first estimated coefficient at 25th QR
scalar firstage=el(firstquartile,1,2)	second estimated coefficient at 25th QR
gen fedu=firstedu*education	role of the first coefficient to lwage at 25th QR
gen fage=firstage*age	role of the second coefficient to lwage at 25th QR

5.A.5 R-codes for the Koenker and Basset test for heteroskedasticity

fit.median = rq(growth~year, method='br')	median regression
summary(fit.median)	
residuals(fit.median)	computes residuals

(continued)

```
taus = c(0.1,0.2,0.3,0.5,0.7,0.8,0.9)                    selected quantiles
fit.deciles = rq(growth~year, method='br',tau=taus)      quantile regressions
out = anova(fit0.10perc, fit0.20perc, fit0.30perc,
            fit0.50perc, fit0.70perc, fit0.80perc,       Koenker and Basset test
            fit0.90perc, test="Wald")
```

Appendix 5.B Dependent data

Stata codes

```
qreg y x                                  median regression
predict res,resid                         save residuals
gen lagres=res[_n-1]                      generates lagged residuals
corr res lagres                           computes correlation coefficient
scalar correlation=r(rho)                 saves the correlation coefficient
gen lagy=y[_n-1]                          generates lagged dependent variable
gen lagx=x[_n-1]                          generates lagged independent variable
qreg y x lagy lagx                        median regression with lagged values
gen ystar=y-correlation*lagy              purge correlation in the y
gen xstar=x-correlation*lagx              purge correlation in the x
qreg ystar xstar                          median regression with transformed variables
```

R codes

```
medianReg = rq(y ~ x, tay=0.5)            median regression
qrResid = residuals(medianReg)            save residuals
corResid = cor(head(qrResid, -1),         computes correlation coefficient
tail(qrResid, -1))                        between residuals and lagged residuals
lagY = tail(y, -1)                        generates lagged dependent variable
lagX = tail(x, -1)                        generates lagged independent variable
medianLagReg = rq(y ~ lagX+lagY+x,        median regression with lagged values
tay=0.5)
summary(medianLagReg)
yStar = head(y) - corResid * lagY         purge correlation in the y
xStar = head(x) - corResid * lagX         purge correlation in the x
medianReg = rq(yStar ~ xStar, tay=0.5)    median regression with transformed
                                          variables
summary(medianReg)
```

6

Additional models

Introduction

The properties of quantile regression (QR) have been described in the previous chapters. Given its properties, QR can be considered a versatile method for use in several frameworks, unlike the classical regression model. This chapter deals with the main extensions of QR: its application in nonparametric models and nonlinear relationships among the variables, in the presence of censored and longitudinal data, when data are derived from different groups and when the dependent variable is dichotomous.

Several real datasets are used to show capabilities of more sophisticated QR models.

6.1 Nonparametric quantile regression

In a linear regression model, the function relating the conditioned values of the response variable to the explanatory variables is a prior known and fixed as a linear function. In real data applications, such a linearity assumption may be too strong and can lead to meaningless results. Nonparametric regression allows the assumption of linearity to be relaxed (Fox 2000b), and it restricts the analysis to smooth and continuous functions.

The aim of nonparametric regression is to identify the best regression function according to the data distribution, rather than to estimate parameters. Let us consider the simplest regression case of one explanatory variable:

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{e}. \tag{6.1}$$

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

In nonparametric regression, the function $f(\cdot)$ is not specified, and it is commonly assumed the errors are normally and independently distributed (*NID*): $\mathbf{e} \sim NID(0, \sigma^2)$. As the final aim is to draw a smooth curve on the data scatter plot, nonparametric regression is also known as scatter smoothing.

To appreciate the philosophy underlying smoothing curves, Scott (1992) noted the requirement of designers of ship hulls to construct smooth curves passing through given points. Their aim was to identify specific points in a flexible, thin and long piece of wood where the minimum strain would be exerted.

In the framework of nonparametric regression, several methods have been proposed in the literature. The present section describes only those methods that have been extended to QR, local polynomial regression and smoothing splines.

6.1.1 Local polynomial regression

A historical review by Cleveland and Loader (1996) revealed that the origins of local regression can be dated back to the nineteenth century. Its modern use was in the time series framework (Macauley 1931), and it was later applied in a regression framework (Watson 1964; Stone 1977; Cleveland 1979). Chaudhuri (1991) was the first to introduce local polynomial QR.

Local polynomial regression (LPR) assumes that $f(\cdot)$ is a polynomial function of a given degree, usually greater than 1. It is straightforward to say that a function of degree 1 corresponds to the classical ordinary least squares (OLS) method. LPR can be considered as an extension of kernel estimation, which is a weighted version of the local averaging. In the following, LPR will be introduced, after some preliminary notes about local averaging regression and kernel estimation.

The mean estimation of the conditioned response variable aims to analyse the behavior of the **y** average value according to the **x** values. The population nonparametric function could be approximated by joining all the conditioned mean values $\bar{y}|x_i$ for each distinct value of the explanatory variable. If the explanatory variable is continuous, it is difficult to find large groups of observations sharing the same value of the explanatory variable. The conditioned mean values become meaningless if they refer to a few observations or even one observation.

Local averaging aims to overcome this drawback by averaging the **y** values in a certain number of intervals (*windows*). The windows are identified by fixing specific **x** values of interest (very often, these comprise all the observations), named *focal values*, and conditioning the window to be centred on the focal values or to contain a constant proportion of observations (*smoother span*), which are the nearest neighbors of the focal values. The fitted **y** values are computed for each **x** focal value. The regression function is represented in the scatter diagram by connecting the points having as coordinates the focal and the fitted values.

The local average produces rough smoothing and therefore it is not able to follow carefully trends in the boundaries.

As the reconstruction of the regression function is realized in different parts of the range of values of the explanatory variable, the procedure is named 'local' with

respect to the classical regression where the relationship between **y** and **x** is analyzed globally, thereby encompassing the whole dataset.

Local averaging can be easily extended to the QR framework if, for each window, the quantile of interest is computed in place of the average.

Kernel estimation, or *local weighted averaging*, is a more sophisticated version of local averaging. This type of estimation is based on the idea of assigning different weights to the observations according to their proximity to the focal values. Thus, smoother results are obtained with respect to local averaging. In *local weighted averaging* regression (Cleveland 1979), two steps must be performed: the distance between each observation and the focal point must be defined and a weight function able to take this distance into account has to be adopted.

Regarding the first step, greater weights are assigned to observations closer to the focal values, where the distance between the i-th observation and the focal value x_0 is the quantity: $z_i = \frac{x_i - x_0}{h}$. The parameter h is named the *bandwidth* and it regulates the degree of smoothness as it is done by the window in the local averaging. Increasing the h parameter results in too much smoothing and hides part of the data structure, but too small a h parameter involves fitting based on small windows, resulting in too much variability (Koenker 2008). Typically, the choice of the smoothing parameter relies on a trade-off between bias and variance (Hastie and Tibshirani 1990). It can be performed using a graphical approach (defining several h values and subjectively choosing the best) or an automatic selection such as one based on cross-validation or Mallow's C_p (Mallows 1973).

The second step consists of finding a weight function, namely a kernel function $K(\mathbf{z})$, that peaks when $\mathbf{z} = 0$ and decays smoothly to 0 as $\mathbf{z}$ increases (Cleveland and Loader 1996). To summarize, the adopted weight function must satisfy the following properties (Cleveland 1979):

1. Weights are positive for observations near the focal value, with $K(\mathbf{z}) > 0$ for $|\mathbf{z}| < 1$.

2. The K function is symmetric because observations on the right-hand side of the focal value and those on the left-hand side must be handled in the same way, with $K(-\mathbf{z}) = K(\mathbf{z})$.

3. Observations near the focal value receive greater weights, with $K(\mathbf{z})$ a nondecreasing function for $\mathbf{z} \geq 0$.

4. Observations far from the focal value receive weights equal to zero for computational reasons, with $K(\mathbf{z}) = 0$ for $|\mathbf{z}| \geq 1$.

The most widely used kernel functions are Gaussian and tricube. Both of these respect the previous properties. The Gaussian kernel function is a standardized normal function where the *bandwidth* h represents the standard deviation. The asymptotic and symmetry properties of the normal function guarantee that observations with a distance from the focal value greater than $2h$ receive weights almost equal to zero.

In the tricube function the *bandwidth* h represents the half-width of a window centred at the focal value:

$$K(\mathbf{z}) = \begin{cases} 1 & for \quad |\mathbf{z}| < 1 \\ 0 & for \quad |\mathbf{z}| \geq 1 \end{cases} \tag{6.2}$$

It has been empirically proved that most of the common weighting functions yield approximately the same results (Venables and Ripley 2002).

For each focal value, the regression function is estimated through a weighted least-squared regression. As in local averaging, kernel smoothers causes high bias in the endpoints (Simonoff 1996).

Koenker *et al.* (1992) proposed a generalization to QR as the solution to the following problem:

$$\min_{[\beta_0, \beta_1(x)] \in R_2} \sum w_i \left\{ y_i - \rho_\theta \left[\beta_0, \beta_1(x_i) \right] \right\}, \tag{6.3}$$

where β_0 is the intercept and β_1 the slope of the explanatory variable.

Equation (6.3) refers to a simple quantile kernel averaging, but it can be easily extended to the multiple case.

Local polynomial regression can be considered an extension of local weighted averaging because a p degree polynomial can be estimated by assigning a weight to each observation:

$$\mathbf{y} = \beta_0 + \beta_1(\mathbf{x} - x_0) + \beta_2(\mathbf{x} - x_0)^2 + \ldots + \beta_p(\mathbf{x} - x_0)^p + \mathbf{e}. \tag{6.4}$$

The generalization to QR is simply obtained by introducing the check function ρ_θ:

$$min_{\beta \in R^{p+1}} \sum_{i=1}^{n} w_i(x) \rho_\theta \left[y_i - \beta_0 - \beta_1(x_i - x_0) - \beta_2(x_i - x_0)^2 - \ldots - \beta_p(x_i - x_0)^p \right]. \tag{6.5}$$

Setting $p = 0$ yields the kernel estimator, while $p = 1$ corresponds to a local linear regression. In the case of a higher order polynomial, it is obviously necessary to take into account powers of the explanatory variables and cross-products between them. For example, if $p = 2$ (local quadratic fitting), the model comprises, beyond the explanatory variables, their squares and their cross-products. The choice of the polynomial degree is derived from a balance between the computational cost and the need to obtain data fitting that is as good as possible. The aim, as in the *bandwidth* choice, is to solve the bias–variance trade–off by adopting a solution able to provide as little bias as possible, but also as little variance as possible. Local linear regression very often represents a good compromise.

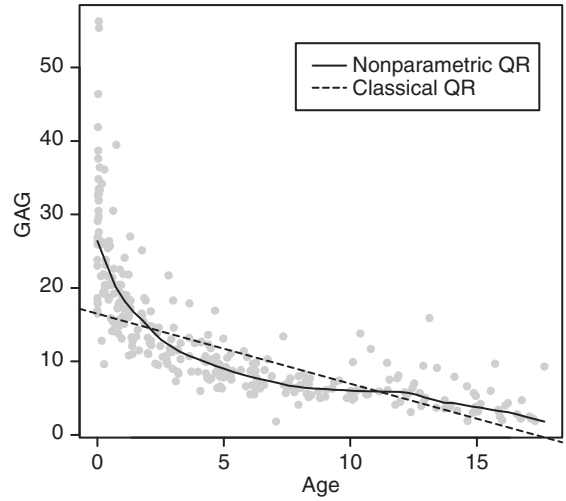

Figure 6.1 Local polynomial regression (solid line) compared with median local polynomial regression (dashed line) with h = *1. The dashed line does not follow the shape of the cloud as well as the median line.*

Classical and quantile nonparametric regressions are compared through the *GAG urine* dataset available in the R package 'MASS' (Venables and Ripley 2002), where concentrations of glycosaminoglycans (GAGs) in urine are analyzed according to age. Dependence analysis is used (Giugliani *et al.* 1990) to detect mucopolysaccharidoses, a family of inheritable disorders caused by a deficiency of GAGs which are lysonal enzymes necessary to degrade mucopolysaccharidoses. Usually the quantity of GAGs decreases with growth. The data comprise 314 observations aged 0 to 17 years. The aim of the analysis is to help predict whether the GAGs concentration (hereafter referred to as *GAG*) can be considered normal according to the patient's age (hereafter referred to as *Age*).

Figure 6.1 shows two curves (the median regression line and the median local polynomial regression with $p = 1$) plotted on the scatter plot of *GAG* and *Age*. The *bandwidth h* has been fixed to 1, and it is evident that a linear fitting would be very inappropriate.

Once a decision has been made to adopt the nonparametric approach, a graphical comparison of results derived from different settings of the *bandwidth* can identify the most suitable choice. In Figure 6.2, the classical approximation to a linear fitting is visualized as the *bandwidth* increases. Looking at the shape of the scatter plot, it seems reasonable to opt for a *bandwidth* as low as possible. Nevertheless, even a *bandwidth* equal to 1 is not able to perfectly depict the extreme parts of the distribution, especially for the lower *Age* values.

Estimating a polynomial curve for a set of quantiles of interest (e.g., 0.1, 0.25, 0.5, 0.75, 0.9) highlights the advantage of using nonparametric QR. In Figure 6.3, each curve describes a part of the distribution of the conditioned dependent variable.

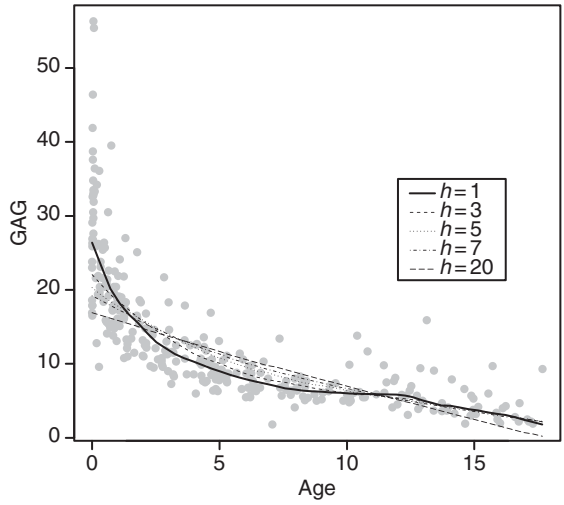

Figure 6.2 Median local polynomial regression lines for different choices of bandwidth. *Increasing the* h *parameter enhances the linear fitting. The fitting able to follow the shape of the cloud corresponds to the lowest* h.

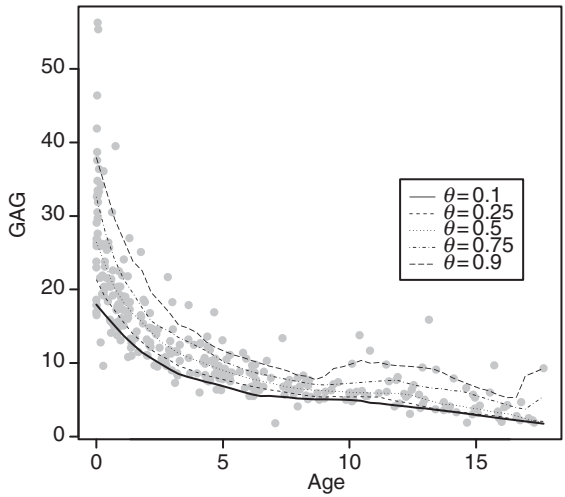

*Figure 6.3 Quantile local polynomial regression lines for different quantiles (*h = *1). Even the extreme parts of the distribution are taken into account. The risk of overfitting due to a low bandwidth is evident for quantile 0.9.*

As the lines in Figure 6.3 have been constructed using a *bandwidth* equal to 1, the extreme parts of the distribution are not well approximated. The drawback related to a low *bandwidth* choice is more evident in the case of $\theta = 0.9$ which shows an overfitting of the data.

6.1.2 Quantile smoothing splines

Splines are a relatively widely used smoothing technique in regression analysis. Briefly, regression splines are obtained joining smoothed polynomial functions separated by a sequence of knots. To obtain a more flexible curve, it is necessary to use more knots. The main issues are the definition of the number of knots and the features of the polynomial functions. In the literature, several data-driven criteria have been proposed to define the number of knots. One of the most commonly used approaches selects knots on the basis of user-defined quantiles of the explanatory variable. Another takes into account regions where $f(x)$ changes more rapidly. The latter approach is aimed at determining whether the first and second derivative of the polynomial functions are constrained to be continuous or discontinuous at the knots. A popular choice is to set both of these as continuous because this enables a smooth curve to be obtained.

A widespread example of regression splines with $f(x)$ having two continuous derivatives is represented by cubic smoothing splines.

A typical smoothing spline is obtained by the introduction of a penalty term in the classical objective function of a regression analysis (Fox 2000a):

$$SS(h) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_{x_{min}}^{x_{max}} \left[f''(x)\right]^2 dx. \tag{6.6}$$

Craig and Ng (2001) defined the previous two terms as *fidelity* and *roughness* (or *penalty*), respectively, where the first measures the goodness of fit of the estimated function and the second represents the degree of smoothness of the estimated fit. The parameter λ plays the same role as the bandwidth in the kernel or in the local polynomial regression. The integral goes from the minimum to the maximum value of the explanatory variable. If λ is set very low, even to zero, the estimated function provides an interpolation of the data. If λ is high, a global linear least-squares fit to the data is obtained because the penalty dominates the fidelity. The parameter λ is connected to the number of interpolated points (let us denote it as p_λ) which can be considered as a measure of the degrees of freedom of the fit (Koenker *et al.* 1992). If λ increases, p_λ decreases and vice versa. In particular, $p_\lambda - 1$ represents the number of fitted segments which are joined to obtain the fitted function. The smoothing parameter balances the fidelity of the data and the smoothness of the fit.

The capability of a smoothing spline to provide linear estimations contrasts with the inability to be generalized to multiple regression.

Quantile smoothing splines have been introduced by Hendricks and Koenker (1992), and several variants have been suggested by Koenker *et al.* (1992). The authors proposed quantile smoothing splines as the minimization of the following expression:

$$min_{f \in C^1} \sum_{i=1}^{n} \rho_\alpha [y_i - f(x_i)] + \lambda \int_0^1 \left(\left|f''(x)\right|^p dx\right)^{1/p}, \tag{6.7}$$

where $p \geq 1$ and there is an appropriately chosen C^1.

Beyond the check function, the main differences between the two formulas (6.6) and (6.7) are the replacement of $\left[f''(x)\right]^2$ with $\left|f''(x)\right|$ and the flexibility in the choice of the parameter p. The authors proved that such changes offer a computational advantage because the linear programming form of the parametric QR can be retained. In particular, $p = 1$ is suggested as the best choice leading to a solution representing a quadratic spline.

An alternative form to express fidelity and penalty was proposed by Koenker *et al.* (1994). They replaced the L_1 penalty on $f''(x)$ with a total variation penalty on $f'(x)$:

$$\sum_{i=1}^{n} \rho_\theta \left[y_i - f(x_i)\right] + \lambda J(f), \tag{6.8}$$

where $J(f)$ is the total variation of the first derivative of f. They also showed that the function f minimizing the quantity (6.8) is a linear spline with knots at the points x_i $(i = 1, \ldots, n)$.

QR smoothing splines are described using the same *GAG urine* dataset. The first step requires the choice of the λ parameter. It can be empirically identified as the value balancing the fidelity and the penalty. For example, in the case of median regression, 10 different λ values (integers from 1 to 10) have been used to estimate the fidelity and the penalty. As can be seen in Figure 6.4, this results in the λ value balancing the two terms at around 6.

The same approach can be applied to other quantiles (0.1, 0.25, 0.75, 0.9) to obtain the best λ values of 3, 2.3, 4.2, and 2.3, respectively (Figure 6.5).

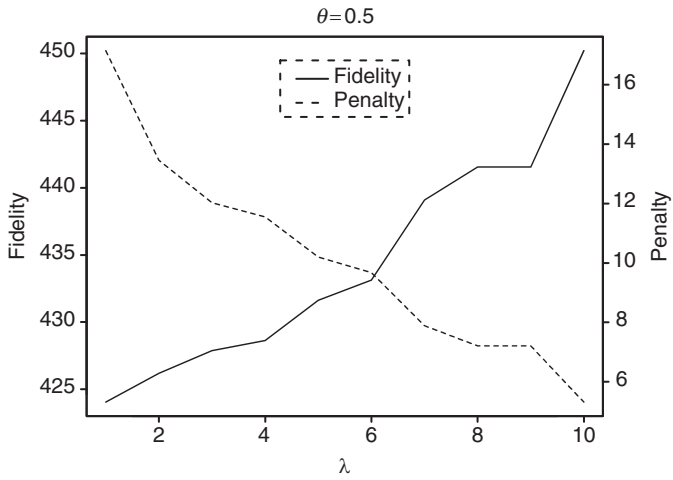

Figure 6.4 Fidelity and penalty values with respect to λ values obtained by a median regression with total variation penalty. Ten different λ values are considered (from 1 to 10 with step equal to 1). A proper choice of λ, balancing fidelity and penalty, is around 6.

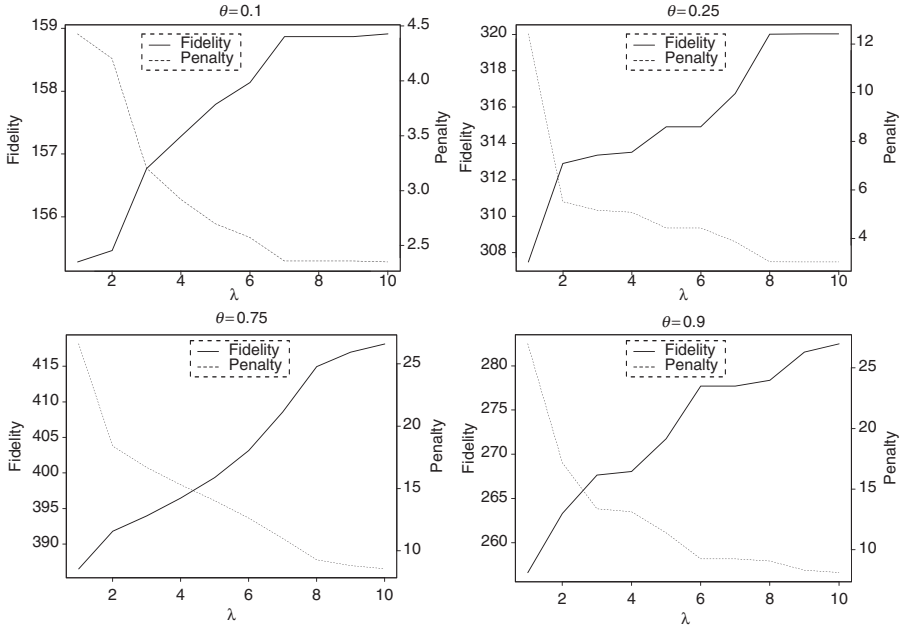

Figure 6.5 Fidelity and penalty values with respect to λ *values obtained by a median regression with the total variation penalty for several quantiles (*θ = *0.1, 0.25, 0.75, 0.9). Ten different* λ *values are considered (from 1 to 10 with step equal to 1). The proper choice of* λ*, balancing the fidelity and the penalty, is 3 (in the case of* θ = *0.1), 2.3 (in the case of* θ = *0.25), 4.2 (in the case of* θ = *0.75) and 2.3 (in the case of* θ = *0.9).*

If the best λ value for each quantile is considered, different cubic splines can be obtained with the total variation method (Figure 6.6). It is important to remember that λ is related to the number of degrees of freedom of the model. Therefore, comparisons among the curves must be made carefully. Each curve in Figure 6.6 describes a part of the distribution. In particular, the two extreme curves corresponding to $\theta = 0.1$ and $\theta = 0.9$ are able to depict parts of the distribution the median regression would fail to consider.

B-splines represent another type of smoothing splines (Hastie and Tibshirani 1990). These differ from classical splines because they use the Bezier curve as a polynomial curve. In Figure 6.7(a), cubic median B-splines curves are shown using a different number of knots: the quartiles or the deciles of the *Age* variable. It is worth noticing that increasing the number of knots causes the curve to become less smoothed.

Assuming a more reasonable choice of a reduced number of knots equal to the quartiles of the *Age* variable, it is possible to appreciate the informative power of different quantile cubic B-splines [Figure 6.7(b)]. As in Figure 6.6, each curve describes a part of the distribution.

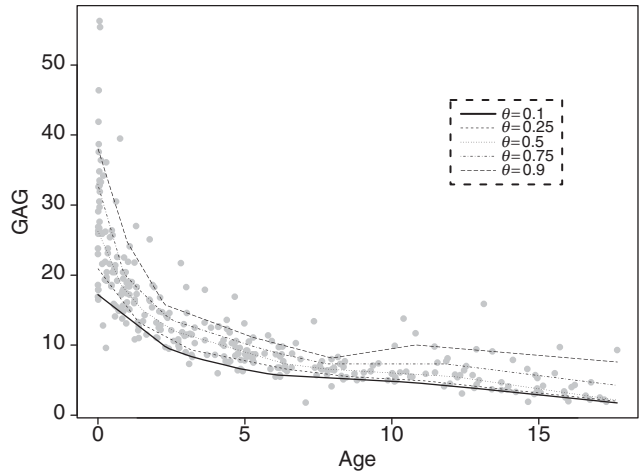

Figure 6.6 QR curves obtained with total variation penalty for several quantiles ($\theta = 0.1, 0.25, 0.5, 0.75, 0.9$) *and* λ *values defined in Figure 6.4 and Figure 6.5.*

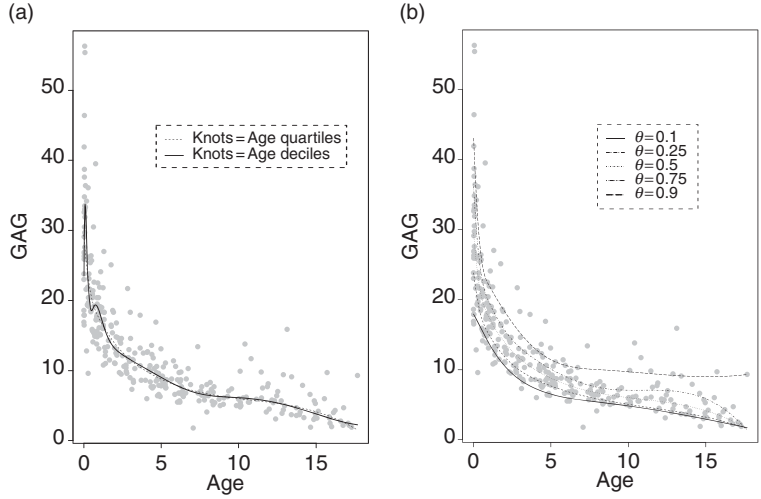

Figure 6.7 Median B-splines for different knots (a). The curve is smoother when Age *quartiles rather than* Age *deciles are used as as knots. In (b), cubic quantile B-splines are shown for different quantiles using* Age *quartiles as knots. Each quantile curve is able to depict a different part of the distribution.*

6.2 Nonlinear quantile regression

The nonlinear approach is very useful as theoretical or empirical issues may prevent a linear approach in many real situations (Gujarati 2004). The term nonlinear refers to the parameters, namely at least one parameter has a power greater than 1, or it is

multiplied or divided by another parameter. In this section, only intrinsically nonlinear models that cannot be made linear with suitable transformations are taken into account.

In a nonlinear model, the function relating the dependent and the explanatory variables is not linear. The functional form of such a relationship is a priori known. Moreover, even if the estimating process is similar to that used in a classical linear regression, iterative procedures are used that require the starting values for the parameters to be specified. Usually, the initial parameter values are based on previous experience or on preliminary analysis (Motulsky and Ransnas 1987).

Inference from a nonlinear regression model cannot follow the same approaches of linear models because the obtained estimators and residuals do not have the same properties as those derived from a classical linear regression analysis. Thus, in nonlinear regression, inference can only be made referring to large sample approximation or to asymptotic theory. In the following, nonlinear regression is treated from only a descriptive point of view.

There is not a great deal of research on nonlinear regression in the QR framework. The most relevant contribution is the interior point algorithm proposed by Koenker and Park (1996). This can be considered an iteratively reweighted least-squares algorithm.

A comparison between nonlinear regression and quantile nonlinear regression is carried out using the *wtloss* dataset available in the R package 'MASS' (Venables and Ripley 2002). The data include a sample of 52 measurements over an 8-month period of the weight of a patient (male, aged 48 years, height 193 cm and a large body frame) participating in a weight-reduction program. As obese patients on weight-reduction programs tend to lose adipose tissue at a diminishing rate, the aim of the analysis is to regress the patient's weight in kilograms (hereafter referred to as *Weight*) according to the time (in days) since the start of the program (hereafter referred to as *Days*).

Venables and Ripley (2002) proposed the following nonlinear regression form to study the dependence of *Weight* (**y**) on *Days* (**x**):

$$\mathbf{y} = \beta_0 + \beta_1 2^{-\mathbf{x}/\beta_2} + \epsilon. \tag{6.9}$$

The first two parameters (β_0 and β_1) are linear, and the third is nonlinear. Each parameter has its own meaning: β_0 is the ultimate lean weight, β_1 is the total weight lost, and β_2 the time taken to lose half of the remaining weight to be lost. Taking into account these interpretations of the parameters, the starting values for the parameters can be identified from the inspection of the scatter plot of *Weight* versus *Days* (Figure 6.8): β_0=90, β_1=95 and β_2=120 .

The estimated curve becomes:

$$\mathbf{y} = 81 + 103 \cdot 2^{-\mathbf{x}/142}, \tag{6.10}$$

and it is superimposed on the scatter plot shown in Figure 6.8.

In the considered example, the nonlinear curve obtained through a least-squares estimation can be viewed as satisfactory. To appreciate the added valued of the QR,

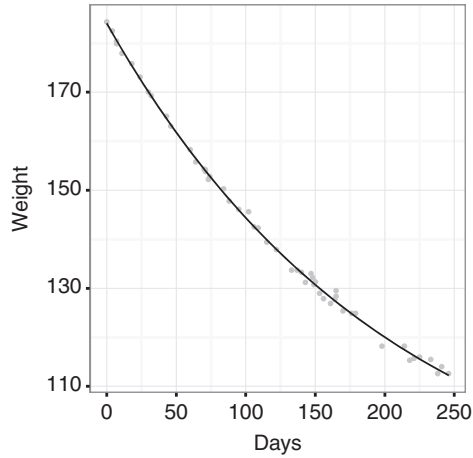

Figure 6.8 The nonlinear curve proposed by Venables and Ripley (2002) superimposed on the scatter plot of Weight *versus* Days*. The curve accurately describes the relationship between the two variables.*

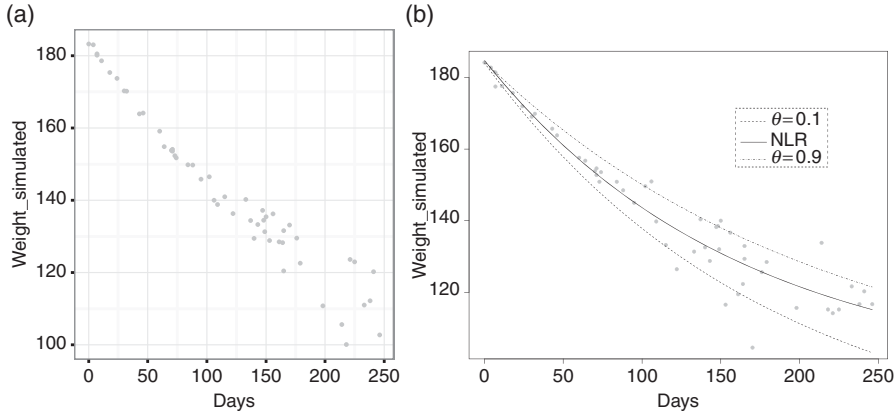

Figure 6.9 Scatter plot of Weight *versus* Days *when more variability is introduced in the* Weight *values (a). In addition to the classical nonlinear curve, two quantile nonlinear curves are plotted (b) corresponding to the quantiles* $\theta = 0.1$ *and* $\theta = 0.9$*. QR complements classical nonlinear regression in the interpretation of the relationship between the two variables.*

the *wtloss* dataset has been disturbed introducing more heteroskedasticity [Figure 6.9(a)]. In such a case, information provided by the classical nonlinear regression curve can be improved using QR. For example, considering two extreme quantiles, $\theta = 0.1$ and $\theta = 0.9$, it is possible to grasp the trend of the dependence relationship in the lower and upper part of the distribution [Figure 6.9(b)] whereas nonlinear regression (solid line) describes only the average trend.

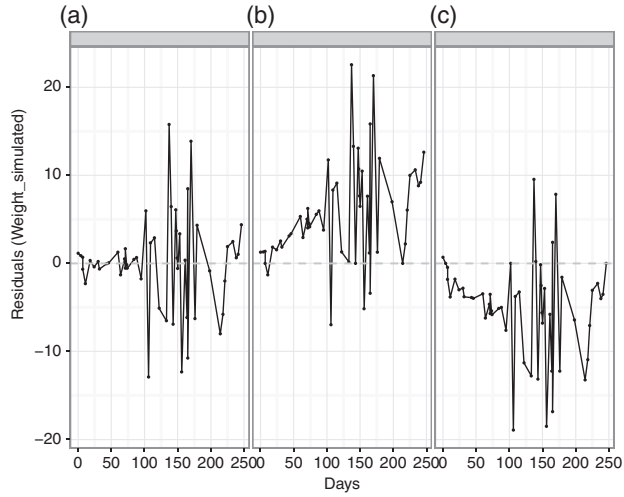

Figure 6.10 Residual plots derived from a classical nonlinear regression (a) and a quantile nonlinear regression with $\theta = 0.1$ *(b) and* $\theta = 0.9$ *(c).*

The different informative contribution of classical and quantile nonlinear regression is also evident by the inspection of the plots of the residuals (Figure 6.10). The classical nonlinear regression represents the average trend of the relationship while losing some nuances both in the lower and the upper part of the distribution. In contrast, nonlinear QR centers on a given part of the distribution, according to the selected quantile.

6.3 Censored quantile regression

QR is a helpful alternative to classical regression when dealing with censored data, which are quite widespread in many fields from biostatistics to econometrics and social sciences. Censored data occur in dependence analysis when values of the explanatory variables are available for all the observations, but values of the dependent variable $\mathbf{y}$ are only known for those observations where $\mathbf{y}$ is greater (left censored) or lower (right censored) than a threshold c. For all the other observations, the only knowledge is that $\mathbf{y} \leq c$ (left censored) or $\mathbf{y} \geq c$ (right censored). In many cases, the censoring threshold is zero. Multiple censoring is also possible if the exact value of the dependent variable is available only for those observations where $c < \mathbf{y} < d$, c and d being two thresholds (Breen 1996).

Censoring models can be also classified as fixed or random. The former is when the thresholds are known for all observations, namely every observation has the same censoring value. The latter refers to cases where the thresholds are observed for only censored observations. An indicator variable takes a value equal to one for uncensored observations and a value of zero for censored ones. Fixed censoring

typically occurs in econometrics, whereas random censoring is mainly used in biostatistics.

In OLS regression, two approaches can be used to handle censored data: an arbitrary value can be assigned to those observations where $\mathbf{y}$ is missing, or $\mathbf{y}$ only can be regressed using the observations where the exact $\mathbf{y}$ value is known. Both the approaches lead to coefficients, which are biased estimates of the population parameters.

To overcome the OLS regression drawbacks previously outlined, Tobin (1958) proposed a model specialized to analyse censored data known as the Tobit model. This model can be expressed as:

$$\mathbf{y}^* = \beta_0 + \beta_1 \mathbf{x} + \mathbf{e}, \tag{6.11}$$

where $\mathbf{y}^*$ is a latent variable containing the true values of the dependent variable for all the observations, $\mathbf{x}$ is the explanatory variable and $\mathbf{e}$ are the errors assumed to be independent and normally distributed with the zero mean and the constant variance. The model can be easily generalized to cases of more than one explanatory variable.

The observed variable $\mathbf{y}$ contains the realizations of $\mathbf{y}^*$ for the uncensored data. In the case of left censoring, it is related to the latent variable as follows:

$$\begin{array}{llll} y_i = y_i^* & if & y_i^* > c & (uncensored\ observations) \\ y_i = c & if & y_i^* \leq c & (censored\ observations), \end{array} \tag{6.12}$$

where i is the i-th observation.

In essence, $\mathbf{y} = max(c, \mathbf{y}^*)$, and the Tobit model in Equation (6.11) can be expressed as:

$$\mathbf{y} = max(c, \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}). \tag{6.13}$$

Right censoring can be similarly modeled by replacing the *max* with the *min*:

$$\mathbf{y} = min(c, \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}) \tag{6.14}$$

because:

$$\begin{array}{llll} y_i = y_i^* & if & y_i^* < c & (uncensored\ observations) \\ y_i = c & if & y_i^* \geq c & (censored\ observations). \end{array} \tag{6.15}$$

Starting from the Tobit model, the censored quantile regression (CQR) model can be derived. In the QR framework, fixed censoring has been proposed by Powell (1986). In the case of random censoring, the most widely used approaches have been proposed by Portnoy (2003) and Peng and Huang (2008).

The fixed censoring model proposed by Powell (1986) is an extension of the least absolute deviations (LAD) estimation method for censored data (Powell 1984). The left censored LAD estimator is the β value that minimizes:

$$S(\beta) = \frac{1}{n} \sum_{i=1}^{n} |y_i - max(0, \beta_0 + \beta_1 x_i)| \tag{6.16}$$

over all β in some parameter space B and considering a threshold equal to zero.

The estimator was proposed as an alternative to the maximum likelihood estimator of the Tobit model and it has been shown to be consistent, asymptotically normal for many distributions and robust to heteroskedasticity (Powell 1984). The censored LAD estimation procedure requires the assumption that the errors are normally distributed.

As in classical QR, the conditional median of the dependent variable does not represent a valid solution in all situations. For example, if many observations are censored at zero, it is advisable to explore quantiles higher than the median to detect the dependence relationship for much of the sample.

The CQR model can be derived by extending the model in Equation (6.16) to other quantiles of interest and taking into account that the $max(0, \beta_0 + \beta_1 \mathbf{x})$ is a monotone transformation of $\mathbf{y}$. As described in Chapter 4, Section 4.1.2, in cases of a monotone transformation of the dependent variable, the quantiles of the transformed $\mathbf{y}$ variable are the transformed quantiles of the original ones.

The CQR model in the case of left censoring becomes:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = max\left[0, \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}\right]. \tag{6.17}$$

As a result, the CQR estimator is obtained by minimizing the quantity:

$$S(\theta) = \sum_{i=1}^{n} \rho_\theta \left\{y_i - max\left[0, \beta_0(\theta) + \beta_1(\theta)x_i\right]\right\} \tag{6.18}$$

over all β in some parameter space $B(\theta)$ and where ρ_θ is the usual check function.

One of the main drawbacks of CQR is the difficulty of estimating many quantiles, from the lowest to the highest. If the number of censored observations is a substantial fraction of the sample, it is difficult to obtain estimates in the upper (in the case of right censoring) or the lower (in the case of left censoring) part of the distribution (Buchinsky 1998).

The minimization of Equation (6.18) is quite complex because it is necessary to minimize a nondifferentiable and nonconvex distance function. Several approaches have been proposed to solve Equation (6.18) (Fitzenberger 1997a). These include the interior point algorithm proposed by Koenker and Park (1996) for nonlinear QR, the iterative linear programming algorithm proposed by Buchinsky (1994) and the widely used BRCENS algorithm proposed by Fitzenberger (1994).

The estimation of the $\beta(\theta)$ parameters requires the assumption that the error distribution is absolutely continuous, with positive density at the θ-th quantile. CQR estimators are consistent and asymptotically normal in large samples [for a demonstration, see Powell (1986) and Fitzenberger (1997b)]. Moreover, the approach proposed by Powell requires that the censoring value is known for all the observations.

To empirically describe the main features of a fixed censored QR model, a simulated dataset is considered with the following characteristics: a sample of $n = 1000$ observations is extracted from the model $\mathbf{y} = 1 + 2\mathbf{x} + (1 + \mathbf{x})\mathbf{e}$ where $\mathbf{x} \sim N(0, 1)$ and

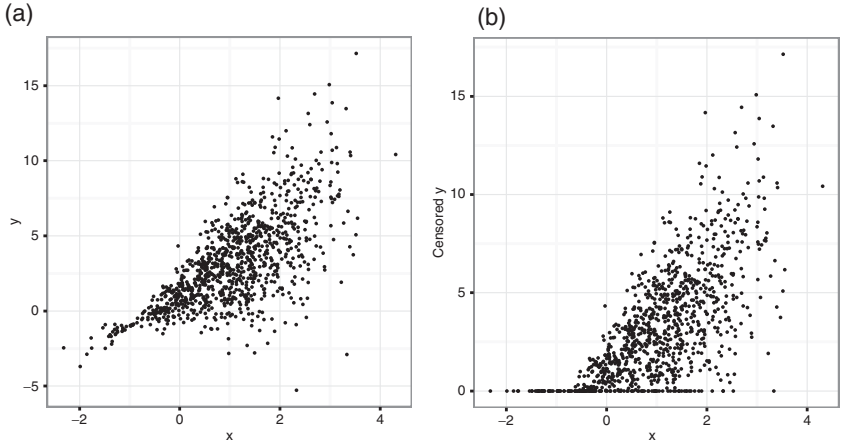

Figure 6.11 Scatter plot of the original (a) and censored (b) simulated data. Values less than zero are considered censored and recoded as zero.

$\mathbf{e} \sim N(0, 1)$. To compare the results obtained with a complete or censored dependent variable, values of the dependent variables less than zero are recoded as zero. Such a procedure makes it possible to simulate a left fixed censored dataset where **y** values equal to zero are considered censored observations. Figure 6.11 shows the scatter plot of the original [Figure 6.11(a)] and censored [Figure 6.11(b)] dependent variable against the **x** variable. Both parts of the figure reveal a heteroskedastic relationship.

Three QR models are estimated and compared using the implementation of Fitzenberger (1994): a classical QR is applied to the original dependent variable and to the censored data, but considering the dependent variable as it was uncensored (the censored observations are treated as zero was their true value). Finally, the Powell algorithm is applied to the censored data (Buchinsky 1998). In the results, figures and tables, the three models are marked, respectively, as *original*, *uncensored* and *censored*. The results derived from the three models estimated on a set of defined quantiles ($\theta = 0.1, 0.25, 0.5, 0.75, 0.9$) are shown in Table 6.1 and Figure 6.12. In all the cases, the coefficients are significantly different from zero, but the values obtained for the censored data are closer to the original ones than those of the uncensored results. The Powell method yields larger estimated standard errors.

Fixed censoring assumes that censoring values are observed for all units. In many practical applications, the Powell estimator is inapplicable because the only available information is whether each observation is censored or not. In such a framework, random censoring is the appropriate solution.

Random censoring typically occurs in survival analysis where the aim is to analyze the time to occurrence of an event during a risk period (period of exposure to a particular risk). A time variable, **t**, representing the time to occurrence of an event, is regressed on a set of covariates through the hazard rate, which measures the risk

Table 6.1 QR results (slopes, standard errors, t-values and p-values) from QR on original, censored and uncensored simulated data for a set of quantiles.

Data	θ	β	Standard error	t-value	p-value
Original	0.10	0.97	0.09	11.07	0.00
	0.25	1.41	0.06	25.68	0.00
	0.50	2.07	0.04	50.54	0.00
	0.75	2.52	0.06	38.84	0.00
	0.90	2.95	0.07	45.06	0.00
Censored	0.10	1.26	0.29	4.31	0.00
	0.25	1.38	0.09	14.87	0.00
	0.50	2.11	0.07	29.39	0.00
	0.75	2.59	0.10	26.18	0.00
	0.90	3.17	0.12	27.43	0.00
Uncensored	0.10	0.60	0.08	7.39	0.00
	0.25	1.24	0.07	16.73	0.00
	0.50	1.89	0.09	21.57	0.00
	0.75	2.24	0.08	27.16	0.00
	0.90	2.51	0.08	30.73	0.00

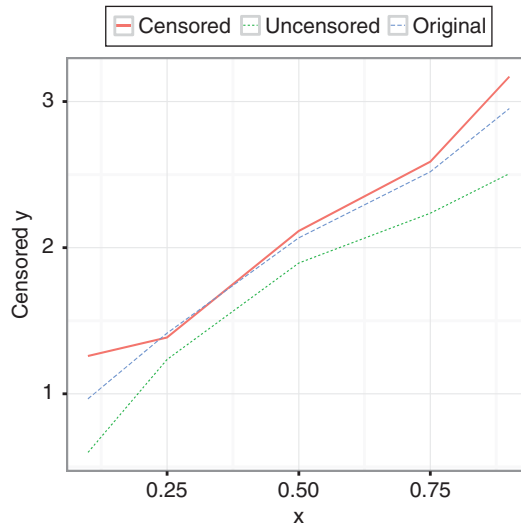

Figure 6.12 Slope coefficients from QR on original, censored and uncensored simulated data. QR applied to the censored data yields coefficients close to those derived from the original data.

of experiencing an event at a given time. The most widely used model in survival analysis is the Cox proportional hazard model (Davison 2003):

$$log\ h(\hat{\mathbf{t}}|\mathbf{X}) = log\ h_0(\hat{\mathbf{t}}) - \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{6.19}$$

where $h(\mathbf{t})$ is the hazard function of the variable $\mathbf{t}$ and h_0 represents a kind of log-baseline hazard because $log\ h(\mathbf{t}|\mathbf{X}) = log\ h_0(\mathbf{t})$ when all covariates are zero. The hazard function measures the risk of experiencing an event at time t, given that the event did not occur before t:

$$h(\mathbf{t}) = lim_{\Delta t \to \infty} \frac{P(t \leq \mathbf{t} < t + \Delta t | \mathbf{t} \geq t)}{\Delta t}. \tag{6.20}$$

The hazard rate can also be expressed as the ratio between the probability density function of $\mathbf{t}$ and the survival function representing the probability of nonoccurrence of an event until time t. There are various types of hazard models, depending on the selected hazard function.

In addition to the Cox proportional hazard model, the accelerated failure time model is another popular model in survival analysis:

$$h(\mathbf{t}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{6.21}$$

A typical choice of the h function is the log transformation.

As both the Cox model and the accelerated failure time model assume that covariates exert a pure location shift in the survival probability, QR can provide a useful support to survival analysis. Koenker and Geling (2001) were the first researchers to formulate QR for survival analysis. Their approach is based on the transformation equivariance property of QR (for details refer to Chapter 4, Section 4.1.2). It is derived from the work of Doksum and Gasko (1990) who traced survival analysis models to the accelerated failure time model in Equation (6.21). QR parameters represent the distance between survival curves related to observations differing by one unit in a single variable and taking fixed values for all the other variables. The proposal of Koenker and Geling (2001) does not hold in cases of censoring data, which are widespread in survival analysis.

Several proposals exist in the literature for estimating QR parameters in cases of random censoring data. Most of the models are right-censored, and the censoring times are assumed to be distributed independently of the regressors and the error terms.

Lindgren (1997) generalized QR to the random right-censored framework using the asymmetric L_1 technique. Conditional quantiles are estimated nonparametrically using the Kaplan–Meier estimator. The main disadvantage with this approach is the computational effort, which makes it suitable for handling only small samples.

Honoré *et al.* (2002) proposed a procedure to estimate a QR random censoring model that involved the adaptation of the fixed censoring approach used by Powell (1984, 1986). Their proposed estimator is consistent under a mild conditional restriction on the error term.

Portnoy (2003) suggested a method to deal with random censoring QR based on the Kaplan–Meier estimate but convergent and asymptotically exact. The proposed algorithm is based on a recursively reweighted estimate of the survival function, which is obtained by redistributing the mass of each censored observation to uncensored ones. The assumption in Portnoy's proposal is that censoring times are independent conditional on covariates (Koenker 2008).

The approach proposed by Portnoy can be considered as complementary to the classical Cox proportional hazard model. Both of them have advantages and disadvantages, depending on the case to be analyzed. Results of the Cox model shed light on hazard rates, whereas QR coefficients measure the effect of the covariates on the survival times. Moreover, QR performs better in cases of heterogeneity of the dependent variable. In addition, the Cox model is able to handle data where covariates are time dependent and in cases of unobserved heterogeneity and competing risks (Fitzenberger and Wilke 2006).

Peng and Huang (2008) proposed another approach to random censoring, extending the martingale representation of the Nelson–Aalen estimator of the cumulative hazard function to the estimation of conditional quantiles. The proposed algorithm is computationally advantageous because it involves minimization only of L_1-type convex functions.

Random censoring QR is described using the *WHAS100* dataset. It contains data gathered by Dr Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School and analyzed in Hosmer *et al.* (2008). The aim of the study is to analyze factors associated with survival rates following hospital admission for acute myocardial infarction[1]. The following variables were studied in a sample of 100 observations: length of the follow-up time in days (hereafter referred to as *lenfol*); follow-up status (hereafter referred to as *fstat*), with categories 1 = *Dead* and 0 = *Alive*; age in years (hereafter referred to as *age*) and body mass index in kg/m^2 (hereafter *BMI*).

The proposed random censoring application considers *lenfol* as the dependent variable, and *fstat* as the indicator variable distinguishing uncensored observations (*Dead*) from censored ones (*Alive*). The BMI is treated as an explanatory variable. Figure 6.13(a) shows the scatter plot of the data according to the *lenfol* and the *BMI*. The algorithm proposed by Portnoy is used to estimate a simple QR model. The trend of the parameters is visualized in Figure 6.13(b) together with 95% confidence intervals. The estimated parameter values and their standard errors and p-values are shown in Table 6.2. The effect of the *BMI* on the *lenfol* increases, moving from the lower to the higher quantiles. Namely, an increase in the *BMI* value has a positive effect on the survival probability, and this effect is higher for individuals in the upper part of the distribution of the dependent variable. It is worth noting that the algorithm was used to estimate several models, but only some of the parameters were estimated because of the presence of censored observations in the upper part of the distribution.

[1] Data are available at ftp://ftp.wiley.com/public/sci_tech_med/survival.

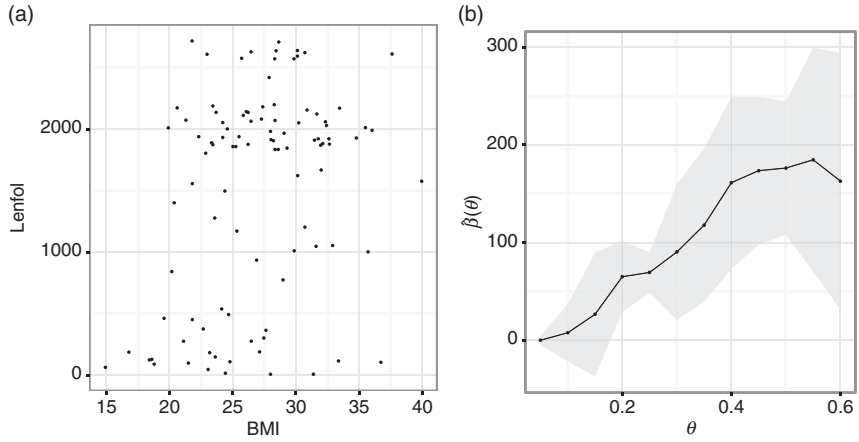

Figure 6.13 Scatter plot of the follow-up time and the BMI *(a) and the slope coefficients of the censored QR model (b).* β *values increase moving from lower to higher quantiles.*

Table 6.2 Censored QR results using the algorithm developed by Portnoy (2003). β values increase moving from lower to higher quantiles.

θ	β	Standard error	p-value
0.05	−0.01	4.83	1.00
0.10	7.64	21.47	0.72
0.15	26.51	39.09	0.50
0.20	65.06	16.79	0.00
0.25	69.42	13.67	0.00
0.30	90.37	36.59	0.01
0.35	117.85	50.61	0.02
0.40	161.33	58.68	0.01
0.45	173.62	43.41	0.00
0.50	176.33	54.76	0.00
0.55	184.71	61.36	0.00
0.60	162.94	59.08	0.01

The QR results can be compared with those derived from a Cox proportional model (Table 6.3). The effect of the *BMI* on the hazard rate is negative. This is consistent with QR results. The effect is significantly different from zero, as shown by the very low p-value. Inference on the slope parameter is provided by a Wald statistic (z in Table 6.3), which is obtained through the ratio of the coefficient (first column) on its standard error (third column). It shows an asymptotically standard normal

Table 6.3 Results of the Cox model. The effect of the *BMI* on the hazard rate is negative and associated with a very low *p*-value.

β	e^{β}	$se(\beta)$	z	p-value
−0.0976	0.907	0.0344	−2.84	0.0045

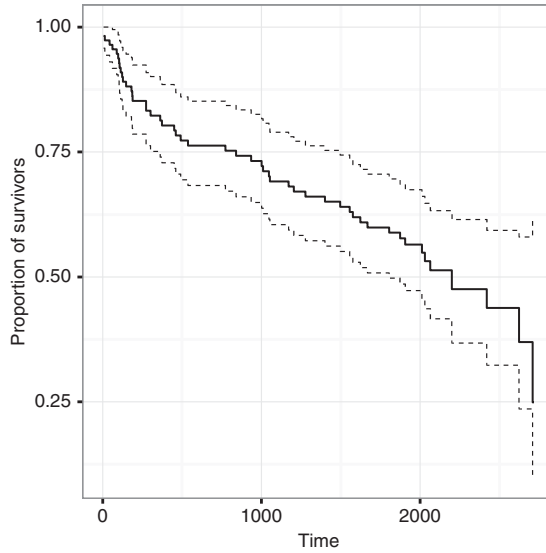

Figure 6.14 Estimated survival function derived from the Cox regression. The broken lines show extremes of a 95% confidence interval around the survival function.

distribution. The results can be also interpreted in terms of a multiplicative effect on the hazard by considering the exponential of the coefficient (second column): an additional value of the *BMI* reduces the daily hazard of not being alive by 9.3%.

The Cox model allows us to estimate the distribution of the survival times by taking the mean values of the covariates. In the *WHAS100* dataset with the *BMI* as explanatory variable, the survival function appears to decrease with time (Figure 6.14), meaning that the proportion of those with a *BMI* equal to the sample mean surviving up to a given time *t*, decreases with time.

6.4 Quantile regression with longitudinal data

Longitudinal data (also known as panel data) refer to repeated measurements of several variables observed on a set of units. The units can be entities of any type (individuals, companies, countries, etc.).

In the statistical literature, there are several methods suited to dealing with longitudinal data such as generalized linear models and generalized linear mixed models. All the approaches restrict the analysis of differences among units in terms of the mean of the dependent variable, and they employ parametric models based on distributional hypothesis. Moreover, methods used to analyze longitudinal data have to take into account that the values observed for each unit are not independent. This fact makes it harder to use inference procedures unless a robust method such as the bootstrap is used (Karlsson 2006).

In the regression framework, longitudinal data are typically structured in a vector of the dependent variable $\mathbf{y}^t_{[n]}$ observed on n units and a matrix of p explanatory variables $\mathbf{X}^t_{[n\times p]}$ where $t = 1, \ldots, T$ is the number of times. Panel data can be balanced if each case is observed for each time occasion and unbalanced if a different number of occasions are observed for each case.

Two of the main approaches to analyze longitudinal data, with fixed or random effects, according to the assumption made about the intercept, are the slope coefficient and the error term (Gujarati 2004).

The aim of a fixed model is to remove the unit time invariant characteristics and to analyze the predictors' net effect. This can be achieved by allowing the intercept of the model to vary across the units but making it constant over time and making the slope coefficients constant across the units and over time. A fixed model with one explanatory variable can be expressed as:

$$\mathbf{y}^t = \boldsymbol{\alpha} + \beta\mathbf{x}^t + \mathbf{e}^t, \tag{6.22}$$

where $\boldsymbol{\alpha}$ is a vector containing the unknown intercept for each unit. It measures the individual effects, for example unobserved heterogeneity, namely, unobserved explanatory variables that are constant over time. As the α values do not vary over time, they are called fixed effects. A classical approach to estimate a fixed-effects model consists of introducing as many variables as the number of units minus one. The coefficient of each dummy variable can be interpreted as the differential intercept fixed coefficient, with the intercept representing the fixed effect of the comparison unit. In Equation (6.22), $\mathbf{e}$ represents the error term, and it is assumed to have the following distribution: $\mathbf{e}^t \sim N(0, \sigma_e^2)$. It is well known that a fixed-effects model is not suited in cases of many units because they affect the number of degrees of freedom of the model and can cause multicollinearity.

The random-effects model aims to overcome this drawback, and it is applied when the variation across units is assumed to be random and uncorrelated with the explanatory variables. A random model can be expressed as:

$$\mathbf{y}^t = \alpha + \beta\mathbf{x}^t + \boldsymbol{\epsilon} + \mathbf{e}^t, \tag{6.23}$$

where α represents the classical average effect, $\boldsymbol{\epsilon}$ is the random deviation of unit intercepts from α and it is distributed as $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2)$.

Most of the literature on QR modeling of longitudinal data centres on fixed-effects models. Combining these models and QR is very fruitful because it facilitates exploration of the effect of the explanatory variables on the whole distribution of the dependent variable, taking into account unobserved heterogeneity.

A QR model for the analysis of longitudinal data with fixed effects can be expressed (Koenker 2004) as:

$$Q_\theta(\mathbf{y}^t|\mathbf{x}) = \boldsymbol{\alpha} + \boldsymbol{\beta}(\theta)\mathbf{x}^t + \mathbf{e}^t. \tag{6.24}$$

This model can be estimated minimizing the following quantity:

$$min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{t=1}^{T} \sum_{i=1}^{n} w_k \rho_\theta \left[y_i^t - \alpha_i - \beta(\theta) x_i^t \right], \tag{6.25}$$

where w_k represents the weights assigned to the quantiles as it is usually performed using L-statistics regression (Mosteller 1946; Koenker 1984). In the quantile fixed-effects model, only the β coefficients change over time.

To appreciate the contribution of QR to the analysis of longitudinal data, the *Panel* dataset is used[2]. It consists of seven countries observed over 10 years (from 1990 to 1999). The aim of the analysis is to study the dependence of a $\mathbf{y}$ variable on a given explanatory variable ($\mathbf{x}_1$), taking into account the effect played by each country over time. The dependent variable displays different behavior over time (Figure 6.15) if considered separately for each country.

Analyzing the distribution of the dependent variable in each country independent of time [Figure 6.16(a)] and in each year independent of the country [Figure 6.16(b)] results in an interesting heterogeneity in both cases.

Let us consider a classical fixed-effects regression and a QR fixed model, with $\theta = 0.5$. The results are compared by means of the estimated $\mathbf{y}$ values. In Figure 6.17, each line represents the trend over time of the estimated $\mathbf{y}$ values with respect to the $\mathbf{x}_1$ variable and for a given country. Differences are apparent particularly for countries A, B and E.

The next step is to verify if such differences increase in cases of more extreme quantiles. Such verification is achieved through a modification proposed by Koenker (2004) of the minimization quantity in Equation (6.25). It is a penalized version able to improve the estimation when the number of units and, consequently, the number of α to be estimated is very high. The minimization function with the penalizing term becomes:

$$min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{t=1}^{T} \sum_{i=1}^{n} w_k \rho_\theta \left[y_i^t - \alpha_i - \beta(\theta) x_i^t \right] + \lambda \sum_{i=1}^{n} |\alpha_i|, \tag{6.26}$$

[2] Data are available at http://dss.princeton.edu/training/Panel101.pdf.

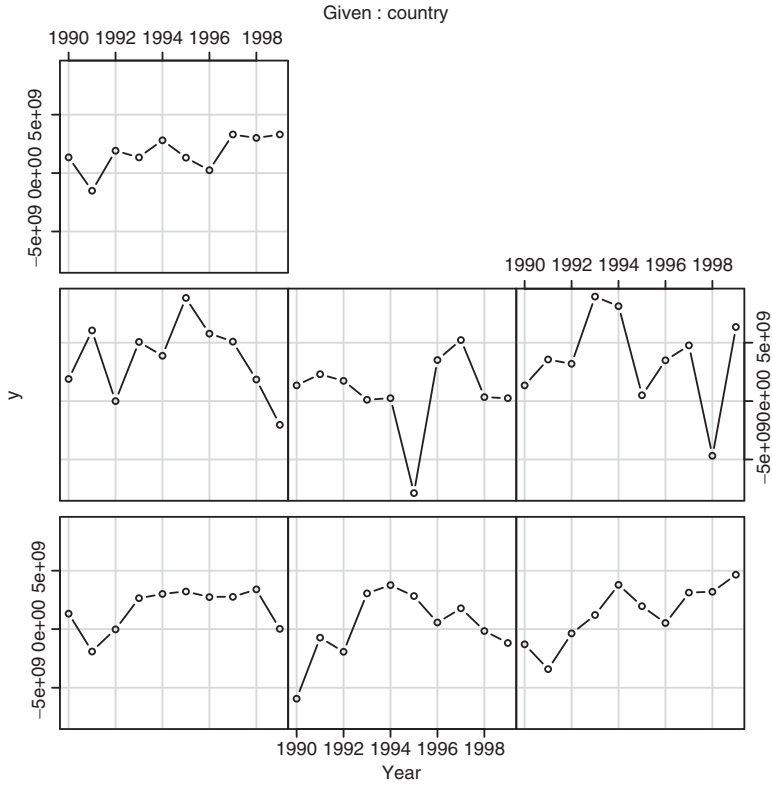

Figure 6.15 Trend of the dependent variable over time and for each country. The horizontal axis represents time and the vertical axis the values of the **y** *variable. Each panel refers to a country. Different trends are shown for each country.*

where λ is the parameter regulating the penalization: λ values towards zero reset to Equation (6.25), and increasing the λ, QR coefficients reveal trends closer to zero.

Considering again a fixed-effects median regression model, coefficient values are plotted in Figure 6.18 according to different values of λ (from 0 to 10) for each country. All the trends tend to zero.

The fixed-effects model in Figure 6.17 reveals a difference mainly for countries A, B and E. Taking into account a wider range of quantiles (from 0.1 to 0.9 with step 0.1), differences also emerge in the other countries (Figure 6.19), and the coefficient trends all increase from lower to higher quantiles.

The approach proposed by Koenker (2004) to analyze longitudinal data through QR has stimulated several contributions on the topic, including the work of Canay (2011), Galvao (2011), Powell (2011), Abrevaya and Dahl (2008) and Geraci and Bottai (2007).

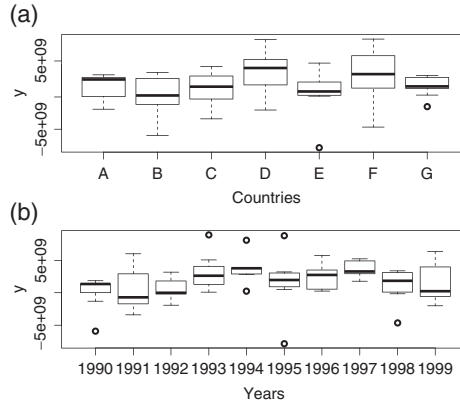

Figure 6.16 Distribution of the dependent variable with respect to countries (a) and to time (b). Both parts of the figure reveal the presence of heterogeneity.

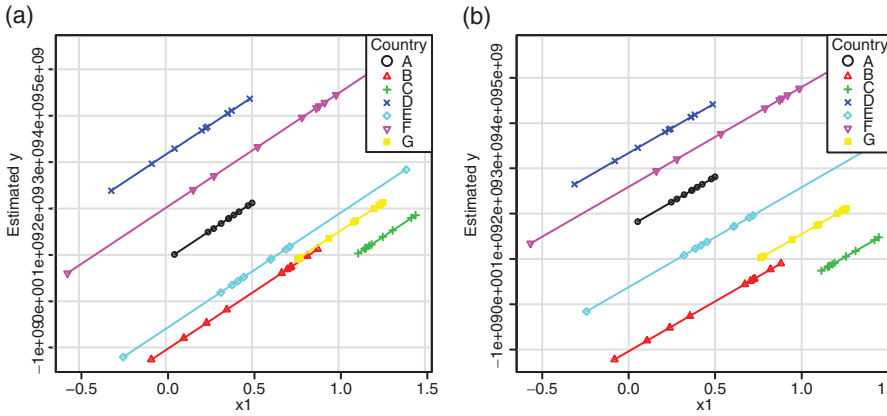

Figure 6.17 Trend of the estimated dependent variable over time with respect to the x_1 variable. Results from a fixed-effects regression (a) and from a fixed-effects median regression (b) are quite similar with the exception of countries A, B and E.

6.5 Group effects through quantile regression

There is widespread research interest in dependence relationships in hierarchical population structures. It is known that if two units have similar features/behaviors or belong to the same group of a stratification variable, the dependence structure of a regression model for the two units is more alike. In such cases, different approaches can be pursued, such as the estimation of different models for each group or the introduction of dummy variables among the regressors denoting group membership. The first approach does not allow us to identify the impact of each group on the dependent variable, and it requires tools for comparisons of the models. The second

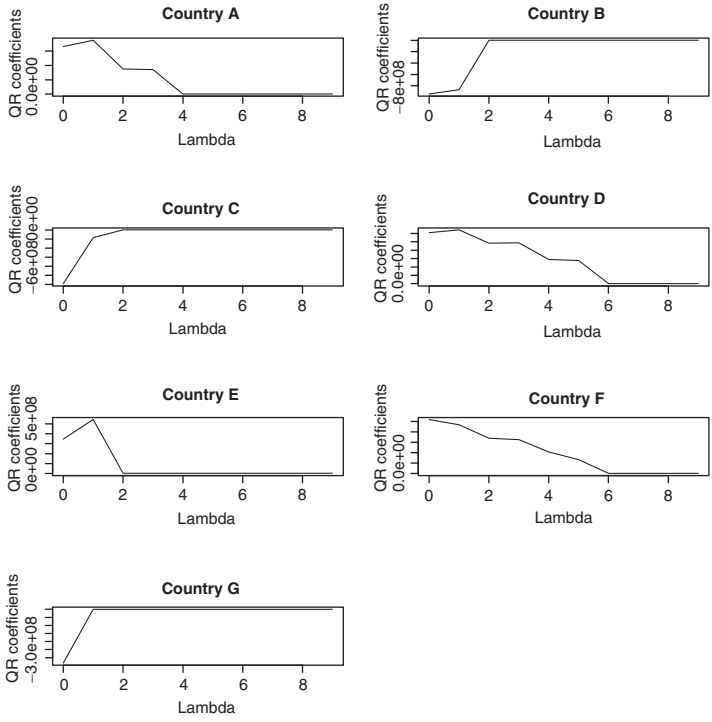

Figure 6.18 Fixed-effects median regression coefficients with respect to λ and according to country. As λ increases, the median coefficients tend to be zero.

approach can detect the effect of each group, but it does not shed light on the impact of the groups on each regressor. Multilevel modeling (Gelman and Hill 2007) is a widely used technique to take into account and explore dependencies in hierarchical population structures. Multilevel modeling restricts the analysis of group differences to the mean of the dependent variable. Such models are parametric and based on distributional hypothesis.

In the QR framework, Davino and Vistocco (2008, 2010) proposed an approach aiming to estimate group effects, taking into account the impact of the covariates on the entire conditional distribution of the dependent variable.

The detection of a typology could derive either from the clustering of units into groups (Davino and Vistocco 2008) or from the analysis of the differences among a priori defined groups. The present section focuses only on the analysis of the differences among groups defined according to a known stratification variable.

Let us consider a data structure composed of a dependent variable vector $\mathbf{y}_{[n]}$ and a matrix $\mathbf{X}_{[n\times p]}$ of regressors. Let the data be row-partitioned in m strata. The generic elements of the response vector and of the regressors matrix are y_i^g and x_{ij}^g (i=1,...,n; j=1,...,p; g=1,...m), where n denotes the number of units, p the number

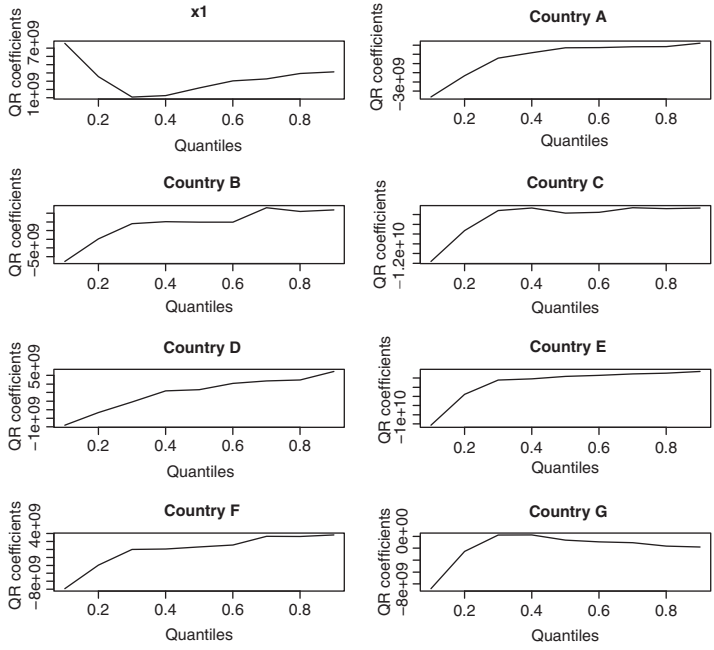

Figure 6.19 Fixed-effects QR coefficients for different θ *values and according to country. The coefficients increase from lower to higher quantiles.*

of regressors and m the number of groups or levels. It follows that n_g is the number of units in group g, and the total sample size can be expressed as $n = \sum_{g=1}^{m} n_g$.

If different models are estimated for each group, the classical linear regression model, for a given group g, can be expressed as :

$$\mathbf{y}^g = \mathbf{X}^g \beta^g + \mathbf{e}^g. \tag{6.27}$$

QR can be viewed as an extension of classical OLS estimation for conditional quantile functions. The QR model for a given conditional quantile θ and for a given group, g, is as follows:

$$Q_\theta^g(\mathbf{y}^g|\mathbf{X}^g) = \mathbf{X}^g \boldsymbol{\beta}^g(\theta) + \mathbf{e}^g. \tag{6.28}$$

QR provides, for each group g, a coefficients matrix, $\hat{\Theta}^g_{[p\times k]}$, whose generic element can be interpreted as the rate of change of the θ-th quantile of the conditional dependent variable distribution per unit change in the value of the jth regressor. The value of k is determined by the number of conditional quantiles that have been estimated.

Starting from the estimated regression quantiles, the density estimation of the response variable can be a useful tool to investigate the effect of a given regressor further (for details, see Chapter 4, Section 4.2). By exploiting the QR estimates, it is

straightforward to estimate the conditional distribution of the response variable for a given group, g, as follows:

$$\hat{\mathbf{y}}^g = \mathbf{X}^g \hat{\boldsymbol{\beta}}^g(\theta). \tag{6.29}$$

The estimated conditional distribution is strictly dependent on the values used for the covariates. It is then possible to use different potential scenarios to evaluate the effect on the conditional response variable, carrying out a what-if study.

If a single analysis is performed for each group, a methodological problem arises because it is necessary to compare m sets of parameters: $\hat{\Theta}^g_{[p\times k]}$, (for $g = 1, \ldots, m$).

The approach proposed by Davino and Vistocco (2008) aims to highlight differences among groups but through a single estimation process.

The approach is structured using the following steps:

1. Global estimation.
2. Identification of the best model for each unit.
3. Identification of the best model for each group.
4. Partial estimation.

In the first step, a QR model is estimated without taking into account the group variable:

$$Q_\theta(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\theta) + \mathbf{e}. \tag{6.30}$$

In the second step, the coefficients matrix $\hat{\Theta}_{[p\times k]}$ and the regressors data matrix $\mathbf{X}$ are used to estimate the conditional distribution matrix of the response variable:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\Theta}. \tag{6.31}$$

The generic element of the $\hat{\mathbf{Y}}_{[n\times k]}$ matrix is the estimate of the response variable in correspondence to the i-th unit according to the θth quantile.

The best model for each unit i is identified by the quantile able to better estimate the response variable, namely through the quantile which minimize the difference between the observed value and the estimated density value:

$$\theta_i : argmin_{\theta=1,\ldots,k} \left| y_i - \hat{y}_i(\theta) \right|. \tag{6.32}$$

The identification of the best quantile for each unit allows the best density estimation vector $\hat{\mathbf{y}}^{best}_\theta$ to be selected from the $\hat{\mathbf{Y}}$ matrix.

In the third step, units are partitioned according to the group variable, and the best model for each group is identified by synthesising the quantiles assigned in the previous step to each unit belonging to the same group: $\hat{\theta}^{best}_g$. The synthesis can be performed through a proper synthesis measure selected according to the particular distribution of $\hat{\mathbf{y}}^{best}_\theta$ in each group. Group effects can be identified analyzing differences among the $\hat{\theta}^{best}_g$.

In the last step, QR is again executed on the total sample. The estimation is carried out for each group on the quantile assigned as the best quantile in the previous step. Differences in the explanatory capability of the regressors according to the group membership can be easily identified through the inspection of a single parameter matrix called $\Theta^{best}_{[p\times m]}$.

The group-effect approach is applied to the *job satisfaction* dataset to measure how the evaluation of features of the job affects overall job satisfaction. Data represent a random sample of interviews with 400 students who had graduated from the University of Macerata and who were working at the time of the interview. The students responded to a questionnaire evaluating different aspects of their university studies and their working lives, such as syllabus, university background, ongoing training, career opportunities, skills, personal interests, free time, salary, office location, job stability, human relationships, stimulating job, independence. Finally, the overall opinion on their job was recorded (hereafter referred to as *job satisfaction*). All the variables were measured on a 10-level scale. As each unit belongs to a group defined by the categories of a stratification variable, it is possible to apply the QR approach previously described for the estimation of group effects. The stratification variable refers to the type of job with categories: self-employed, private employee, or public employee.

The main descriptive statistics of the overall evaluation of job satisfaction (minimum = 1, Q1 = 7.00, mean = 7.68, median = 8.00, Q3 = 8.85, maximum = 10) show the presence of skewness in the distribution.

In Figure 6.20, OLS and QR coefficients are graphically represented for the different evaluations of job features. The horizontal axis displays the different quantiles and the quantile coefficients are represented on the vertical axis. The dashed lines parallel to the horizontal axis correspond to OLS coefficients. OLS coefficients measure the change in the conditional mean, and the QR coefficients measure the change in a given conditional quantile.

The graphical representation aids visualization of the different effects of the evaluation of the features of the job on overall job satisfaction. All the coefficients related to fulfillment aspects (career opportunities, skills, personal interest, free time), increase, moving from the lower to the upper quantiles, whereas the coefficients related to tangible aspects (salary, office location, job stability) move in the opposite direction, with the exception of the office location.

Figure 6.20 can be useful to explore, from a descriptive point of view, the trends in the coefficients along the different quantiles and in comparison with the OLS ones. To take into account the generalization capability of the model, the coefficients estimated by OLS and QR (the following quantiles are considered: 0.1, 0.25, 0.5, 0.75 and 0.9) are shown in Table 6.4 with significant coefficients at $\alpha = 0.1$ highlighted in bold.

Classical OLS estimates provide limited information about the contribution of features of the job to overall job satisfaction compared with the results of QR, which widens the set of significant coefficients and yields a detailed description of the factors influencing the whole conditional distribution of the job satisfaction.

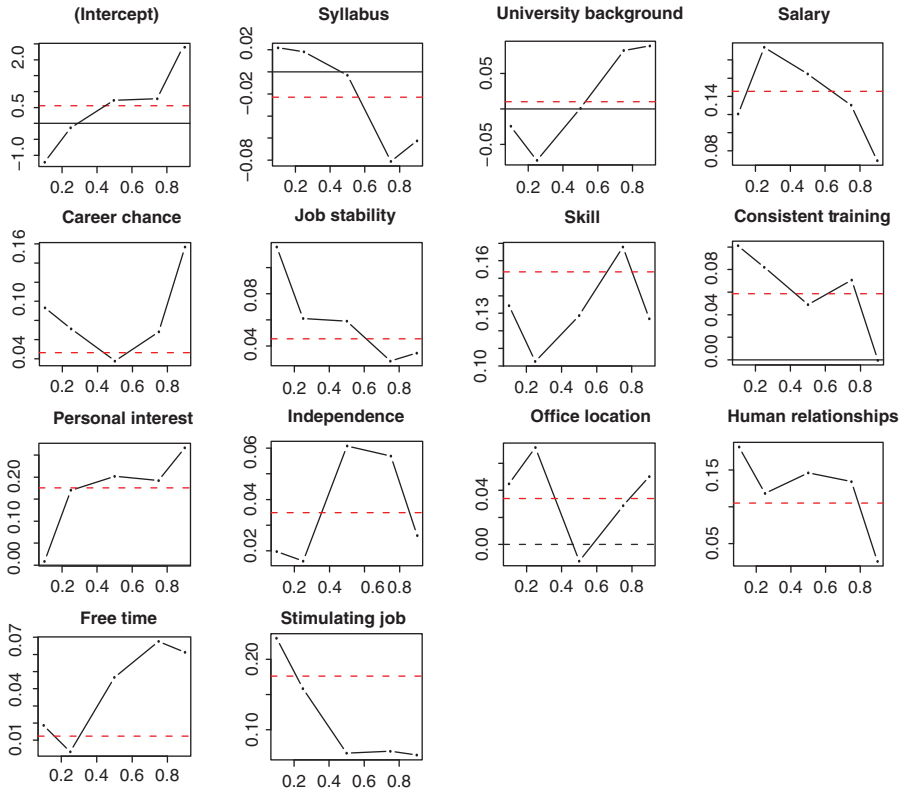

Figure 6.20 OLS and QR coefficients. OLS and QR coefficients are graphically represented for the different evaluations of job features. The horizontal axis displays the quantiles while the quantile coefficients are represented on the vertical axis. The dashed lines parallel to the horizontal axis correspond to OLS coefficients.

The second step requires the identification of the best model for each unit. This is based on the quantile best able to estimate the response variable. The added value in considering the reconstructed estimated response variable $\hat{\mathbf{y}}_{\theta}^{best}$ instead of an OLS estimated response variable is evident form Figure 6.21, which reproduces the histograms of the dependent variable and the estimated dependent variable using OLS or the proposed QR approach.

The results in Table 6.4 and Figure 6.20 refer to the whole sample. They provide global estimations without considering the type of job. In the third step of the procedure, the units are partitioned according to the type of job. Figure 6.22 shows the distribution of the 'best' quantiles assigned to units, grouped according to the type of job. The three distributions appear different, a sign that there is variability among groups. The best model for each category of the grouping variable is identified through the mean value of the 'best' quantiles assigned to units belonging to the gth group: $\theta_1^{best} = 0.371$; $\theta_2^{best} = 0.474$; $\theta_3^{best} = 0.548$.

Table 6.4 OLS and QR coefficients. Significant coefficients at $\alpha = 0.1$ are highlighted in bold.

Variable	LS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
Intercept	0.403	**−1.211**	−0.149	0.711	0.761	**2.370**
Syllabus	−0.009	0.022	0.018	−0.003	−0.081	−0.062
University background	0.004	−0.024	−0.072	0.001	0.082	0.089
Salary	**0.146**	**0.120**	**0.194**	**0.165**	**0.130**	0.069
Career opportunities	**0.078**	0.093	0.071	0.037	0.068	**0.157**
Job stability	**0.061**	**0.116**	0.061	0.059	0.028	0.035
Skills	**0.117**	0.134	0.102	**0.129**	**0.168**	0.127
Ongoing training	0.043	0.101	0.082	0.049	0.070	−0.000
Personal interest	**0.187**	0.008	**0.170**	**0.202**	**0.192**	**0.267**
Independence	0.051	0.019	0.016	0.061	0.056	0.026
Office location	0.031	0.044	0.072	−0.012	0.029	0.050
Human relationships	0.126	**0.181**	0.118	**0.146**	**0.134**	0.026
Free time	0.017	0.189	0.003	0.047	**0.067**	0.061
Amusing job	**0.147**	**0.230**	**0.158**	0.066	0.069	0.064

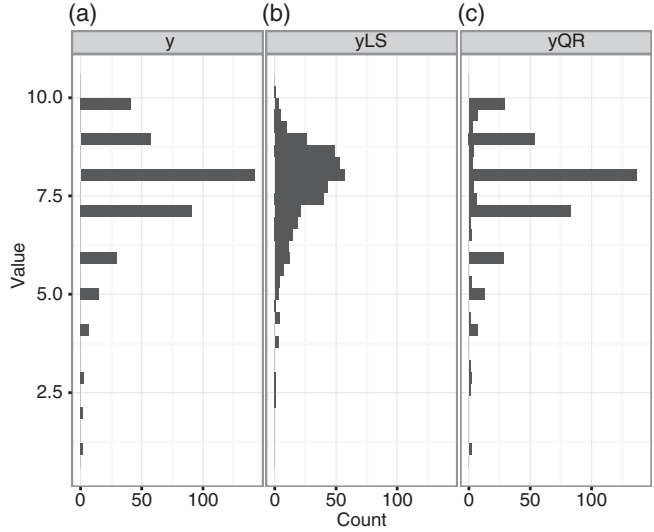

Figure 6.21 Distribution of the dependent variable (a) and of the estimated dependent variable using OLS (b) or the proposed QR approach (c). The estimated response variable is able to reconstruct the original **y** *variable better than the OLS estimated response variable.*

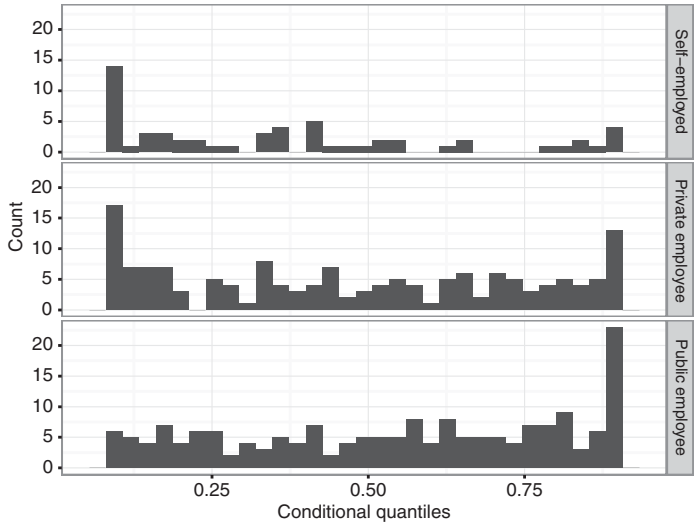

Figure 6.22 Distribution of the 'best' quantiles according to the type of job. The three groups show different distributions.

Table 6.5 QR coefficients with group effects.

Variable	Self-employed	Private employee	public employee
Intercept	0.646	0.683	0.694
Syllabus	−0.007	−0.012	−0.035
University background	−0.030	0.006	0.026
Salary	**0.201**	**0.152**	**0.160**
Career chance	0.012	0.037	−0.008
Job stability	0.049	0.034	0.054
Skill	**0.118**	**0.156**	**0.184**
Consistent training	0.065	0.066	0.064
Personal interest	**0.200**	**0.175**	**0.202**
Independence	0.022	0.035	0.035
Office location	0.011	−0.006	0.007
Human relationships	**0.114**	**0.152**	**0.107**
Free time	0.018	0.032	0.026
Stimulating job	**0.148**	**0.124**	**0.141**

Finally, QR is again executed on the total sample, but only the quantiles assigned to each category of the grouping variable are retained. In Table 6.5, the values of the final parameters matrix $\hat{\Theta}^{best}_{[p\times m]}$ are shown. Significant coefficients at $\alpha = 0.1$ are shown in bold for each covariate (row of the table) and for each type of job (column of the table).

QR coefficients with group effects reveal slight differences among the groups if regressors are ranked according to their coefficients. Notwithstanding this factor, differences among the groups can be identified in terms of the intensity of the values. For example, the evaluation of salary and fulfilling job has a major impact on the overall satisfaction in the self-employed group. Private employees are mainly differentiated from the other groups by the effect of the human relationships variable. Finally, the highest coefficients in the public employees groups are in skill and personal interest.

The approach described in this section can be extended to cluster units according to the dependence structure without a priori information but only using the observed similarities among them in terms of conditional quantile estimates (Davino and Vistocco 2008).

6.6 Binary quantile regression

Models for binary data are quite widespread in many fields of application. For example, they allow selection processes, the realization of events or an attribute presence to be described. The aim of binary models is to evaluate how much the explanatory variables impact on the probability that the event of interest occurs (namely that the dependent variable assumes a value equal to 1). For example, such a probability can represent the propensity to make a choice, the risk of an event or the possibility to gain an attribute.

A probability model assumes the following form:

$$P(\mathbf{y} = 1|\mathbf{X}) = f(\mathbf{X}\beta), \tag{6.33}$$

where f is a known distribution function.

In the regression framework, the most widespread approaches to dealing with binary response data are the linear probability model, the logit model and the probit one. Inferential problems and drawbacks related to the choice of the most suited functional form prevent their use in all situations. In the linear probability model, the error distribution is highly variable and binomial. Moreover, using a linear relationship between the dependent variable and the explanatory variables can be too simplistic and the estimated values are not assured to be in the range $[0, 1]$. The logit model and the probit model introduce a nonlinear relationship with, respectively, a logistic and normal functional form. They provide estimated values in the range of the dependent variable values. Typical nonparametric alternatives to the regression models can be neural networks and classification trees.

QR can represent a valid alternative to the above models when the exploration of the full conditioned distribution of the dependent variable is of great interest. In many real data applications, the extreme quantiles of the response distribution can be of utmost importance and should not be overlooked (Miguis *et al.* 2012).

The origin of binary QR can be dated back to Manski (1975, 1985). He proposed that the maximum score estimator, equivalent to the binary QR estimator, was consistent but not asymptotically normal. Horowitz (1992) proposed an

asymptotically normal distributed and smoothed median estimator, and Kordas (2000, 2006) extended this to general quantiles.

Let us consider a classical regression model:

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{6.34}$$

where $\mathbf{y}^*$ is a continuous latent variable, $\mathbf{X}$ a matrix of explanatory variables and $\mathbf{e}$ is the error component.

The observed dependent variable $\mathbf{y}$ is derived from the following monotone transformation of $\mathbf{y}^*$:

$$\mathbf{y} = I(\mathbf{y}^* \geq 0), \tag{6.35}$$

where I is the indicator function.

In the case of a QR, as the distribution of $\mathbf{e}$ at the θ-th quantile equals 0, Equation (6.34) corresponds to:

$$Q_\theta(\mathbf{y}^*|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\theta) + \mathbf{e}, \tag{6.36}$$

and

$$P\left[\mathbf{y}^* \geq \mathbf{X}\boldsymbol{\beta}(\theta)|\mathbf{X}\right] = 1 - \theta. \tag{6.37}$$

The QR model in terms of the observed variable $\mathbf{y}$ can be easily derived due to the equivariance property of QR in relation to monotone transformations of the dependent variable (see Chapter 4, Section 4.1.2):

$$Q_\theta(\mathbf{y}|\mathbf{X}) = I\left[\mathbf{X}\boldsymbol{\beta}(\theta) \geq 0\right]. \tag{6.38}$$

The β vector is selected by minimizing the following quantity:

$$min_{[\beta:\|\beta\|=1]} \sum_{i=1}^{n} \rho_\theta \left\{y_i - I\left[x_i\beta(\theta) \geq 0\right]\right\}. \tag{6.39}$$

To identify the estimator that respects the binary nature of the data, the condition $\|\beta\| = 1$ is imposed. An alternative can be to fix the value of one parameter equal to 1. The condition on the parameters can appear strong if no a priori information is known, but it is particularly useful for the comparisons among coefficients across models. Moreover, it is reasonable from a numerical point of view because parameters in Equation (6.38) are not affected by a scale change:

$$I\left[\alpha\mathbf{X}\boldsymbol{\beta}(\theta) \geq 0\right] = I\left[\mathbf{X}\boldsymbol{\beta}(\theta) \geq 0\right] \tag{6.40}$$

for all $\alpha > 0$ (Koenker 2005).

From the relationship between $\mathbf{y}$ and $\mathbf{y}^*$ ($\mathbf{y} = I(\mathbf{y}* \geq 0)$), the corresponding probability estimates are derived:

$$P(\mathbf{y} = 1|\mathbf{X}) = P(\mathbf{y}^* \geq 0|\mathbf{X}). \tag{6.41}$$

From Equation (6.37) and Equation (6.41) it is possible to obtain all the possible solutions for the posterior probability of the response, giving the covariates:

$$P\left[\mathbf{y} = 1|\mathbf{X}\boldsymbol{\beta}(\theta) \gtreqless 0\right] \gtreqless (1 - \theta). \tag{6.42}$$

6.7 Summary of key points

- QR local averaging is obtained by computing the dependent variable quantile of interest in place of the average for each defined interval value of the explanatory variable.
- Local weighted averaging and local polynomial regression can be generalized to QR by introducing the check function in the objective function.
- QR provides valuable support in cases of intrinsically nonlinear models.
- In survival analysis, QR coefficients measure the effect of the covariates on the survival times. QR performs better than classical methods such as the Cox model, in cases of heterogeneity of the dependent variable.
- In longitudinal data analysis, QR allows the effect of the explanatory variables on the whole distribution of the dependent variable to be explored, taking into account unobserved heterogeneity.
- If units belong to different groups, QR is able to estimate group effects, taking into account the impact of the covariates on the entire distribution of the dependent variable.

References

Abrevaya J and Dahl CM 2008 The effects of birth inputs on birth weight: evidence from quantile estimation on panel data. *Journal of Business and Economic Statistics* **26**, 379–397.

Breen R 1996 *Regression Models. Censored, Sample Selected or Truncated Data.* Series: Quantitative Applications in the Social Sciences, Volume 111. SAGE Publications, Inc

Buchinsky M 1994 Changes in the U.S. wage structure 1963–1987: application of quantile regression. *Econometrica* **62**, 405–458.

Buchinsky M 1998 Recent advances in quantile regression models: a practical guideline for empirical research. *The Journal of Human Resources* **33**(1), 88–126.

Canay IA 2011 A simple approach to quantile regression for panel data. *Econometrics Journal* **14**, 368–386.

Chaudhuri P 1991 Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis* **39**, 246–269.

Cleveland WS 1979 Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cleveland WS and Loader CL 1996 Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing* (Haerdle W and Schimek M G eds), pp. 10–49. Springer.

Craig SG and Ng PT 2001 Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *Journal of Urban Economics* **49**(1), 100–120.

Davino C and Vistocco D 2008 Quantile regression for the evaluation of student satisfaction. *Italian Journal of Applied Statistics* **20**, 179–196.

Davino C and Vistocco D 2010 Quantile regression for group effect analysis. In *e- book of Compstat2010*. pp. 911–918. Physica-Verlag.

Davison AC 2003 *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Doksum K and Gasko M 1990 On a correspondence between models in binary regression and survival analysis. *International Statistical Review* **58**, 243–252.

Fitzenberger B 1994 *A Note on Estimating Censored Quantile Egressions.* Center for International Labor Economics, Discussion Paper No. 14. University of Konstanz.

Fitzenberger B 1997a Computational aspects of censored quantile regression. *Lecture Notes-Monograph Series* **31**, 171–186.

Fitzenberger B 1997b A guide to censored QR. In *Handbook of Statistics 15: Robust Inference* (Maddala GS and Rao CR eds), Volume 15, pp. 405–437. North-Holland.

Fitzenberger B and Wilke RAA 2006 Using quantile regression for duration analysis. *Allgemeines Statistisches Archiv* **90**, 105–120.

Fox J 2000a *Nonparametric Simple Regression.* Series: Quantitative Applications in the Social Sciences, Volume 130. SAGE Publications, Inc.

Fox J 2000b *Multiple and Generalized Nonparametric Regression.* Series: Quantitative Applications in the Social Sciences, Volume 131. SAGE Publications, Inc.

Galvao JAF 2011 Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics* **164**(1), 142–157.

Gelman A and Hill J 2007 *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Geraci M and Bottai M 2007 Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**, 140–154.

Giugliani R, Dutra-Filfo CS, Barth ML, Enk V and Netto CA 1990 Age-related concentrations of glycosaminoglycans in random urine: a contribution to the laboratorial detection of mucopolysaccharidoses. *Revista Brasileira de Genetica* **13**(3), 599–605.

Gujarati DN 2003 *Basic Econometrics*. McGraw-Hill, International Edition.

Hastie TJ and Tibshirani RJ 1990 *Generalized Additive Models.* Monographs on Statistics & Applied Probability 43 Chapman & Hall.

Hendricks W and Koenker R 1992 Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association* **87**(417), 58–68.

Honoré B, Khanb S and Powell JL 2002 Quantile regression under random censoring. *Journal of Econometrics* **109**, 67–105.

Horowitz JL 1992 A smoothed maximum score estimator for the binary response model. *Econometrica* **60**, 505–531.

Hosmer DW, Lemeshow S and May S 2008 *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd Edition. John Wiley & Sons, Ltd.

Karlsson A 2006 *Estimation and Inference for Quantile Regression of Longitudinal Data with Applications in Biostatistics.* Acya Universitatis Upsaliensis. Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. Uppsala.

Koenker R 1984 A note on L-estimators for linear models. *Statistics and Probability Letters* **2**, 323–325.

Koenker R 2004 Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**(1), 74–89.

Koenker R. 2005 *Quantile Regression*. Cambridge University Press.

Koenker R 2008 Censored quantile regression redux. *Journal of Statistical Software* **27**(6), 1–25.

Koenker R and Geling R 2001 Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association* **96**(454), 458–468.

Koenker R, Ng P and Portnoy S 1994 Quantile smoothing splines. *Biometrika* **81**(4), 673–680.

Koenker R and Park BJ 1996 An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics* **71**, 265–283.

Koenker R, Portnoy S and Ng P 1992 Nonparametric estimation of conditional quantile functions. In *Statistical Data Analysis Based on the L1 Norm and Related Methods* (Dodge Y ed.), pp. 217–229. Elsevier Science.

Kordas G 2000 *Binary Regression Quantiles.* PhD Dissertation, University of Illinois, Urbana-Champaign.

Kordas G 2002 Credit scoring using binary quantile regression. In *Statistical Data Analysis Based on the L1 Norm and Related Methods* (Dodge Y ed.), pp. 125–137. Elsevier Science.

Kordas G 2006 Smoothed binary regression quantiles. *Journal of Applied Economics* **21**, 387–407.

Lindgren A 1997 Quantile regression with censored data using generalized L_1 minimization. *Computational Statistics & Data Analysis* **23**, 509–524.

Macauley FR 1931 *The Smoothing of Time Series*. National Bureau of Economic Research.

Mallows CL 1973 Some comments on C_P. *Technometrics* **14**(4), 661–675.

Manski CF 1975 Maximum score estimation of the stochastic utility. *Journal of Econometrics* **3**, 205–228.

Manski CF 1985 Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**, 313–333.

Miguis VL, Benoit DF and Van den Poel D 2012 *Enhanced Decision Support in Credit Scoring Using Bayesian Binary Quantile Regression.* Working paper, Faculty of Economics and Business Administration, Ghent University, D/2012/7012/36.

Mosteller F 1946 On some useful inefficient statistics. *Annals of Mathematical Statistics* **17**, 377–408.

Motulsky HJ and Ransnas LA 1987 Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *The Journal of the Federation of America Societies for Experimental Biology* **1**(5), 365–74.

Peng L and Huang Y 2008 Survival analysis with quantile regression models. *Journal of the American Statistical Association*, **103**(482), 637–649.

Portnoy S 2003 Censored regression quantiles. *Journal of the American Statistical Association* **98**(464), 1001–1012.

Powell D 2011 *Unconditional Quantile Regression for Panel Data with Exogenous or Endogenous Regressors.* Working paper, RAND Corporation.

Powell JL 1984 Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* **25**, 303–325.

Powell JL 1986 Censored regression quantiles. *Journal of Econometrics* **32**, 143–155.

Scott DW 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd.

Simonoff JS 1996 *Smoothing Methods in Statistics.* Springer Series in Statistics Springer.

Stone CJ 1977 Consistent nonparametric regression. *The Annals of Statistics* **5**, 595–620.

Tobin J 1958 Estimation of relationships for limited dependent variables. *Econometrica* **26**(1), 24–36.

Venables WN and Ripley BD 2002 *Modern Applied Statistics with S*, 4th Edition. Springer.

Watson GS 1964 Smooth regression analysis. *Sankhya, Series A*, **26**, 359–372.

Appendix A

Quantile regression and surroundings using R

Introduction

The features of the software R in the QR framework are described using a hypothetical data set, hereafter referred to as *example.{txt, csv, xls, dta, sav}* where the file extensions are associated with the following file formats:

txt tab delimited text file;
csv comma separated values file;
xls spreadsheet (Microsoft Excel) file;
dta Stata file;
sav Spss data file.

In the following we assume that the file contains a dependent numerical variable (y), a set of numerical explanatory variables ($x1, x2, x3, x4$), and a set of categorical explanatory variables ($z1, z2$).

This appendix shows the R commands useful to exploit a data analysis, starting with the data importation until the regression analysis. R commands are shown in bold font, comments using regular font. Of course, we do not claim to be exhaustive. There are several references for learning the R language. The main starting point is the official documentation. The commands shown in this appendix are fully explained in the introductory R manual (R Core Team 2012b; R Core Team 2012c). There are several books on the R system; Dalgaard (2008) is the natural extension

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

of the official base documentation. Among the myriad of books available on the subject, we refer the interested reader to Cohen and Cohen (2008), Teetor (2011) and Venables and Ripley (2002).

A.1 Loading data

A typical situation involves the importation of data starting from a text file. In particular the two more common formats are tab delimited text (txt) files and comma separated values (csv) files. We assume the file is in the current working directory.

```
#to obtain the current working directory
getwd()

#to set the working directory by specifying the complete
#path (in the example the working directory is set to
#the subfolder bookQR cointained in the folder
#Documents)
setwd("/Documents/bookQR")
```

A.1.1 Text data

The main format used for storing text data is the delimited file. The particular character used as column delimiter differs among the files, the two most commonly used being the *tab* character and the *comma* character.

```
#to import data from a tab delimited text file
#through the general read.table function
dataExample <- read.table("example.txt", header=T,
                          sep="\t")
#it is possible to exploit the function read.delim
#to import data from a tab delimited text file
#where the first row contains the columns' headers
dataExample <- read.delim("example.txt")
#NOTE: the read.delim2 file is a variant for countries
#that use a comma as decimal point and a semicolon
#as field separator to import data from a comma
#separated text file through general read.table function
dataExample <- read.table("example.txt", header=T,
                          sep=",")
```

```
#to import data from a csv file where the first row
#contains the columns' headers, it is possible to
#exploit the function read.csv
dataExample <- read.csv("example.txt")
#NOTE: the read.csv2 function is a variant for
#countries that use a comma as decimal point and a
#semicolon as field separator
```

A.1.2 Spreadsheet data

An equally widespread file format is the spreadsheet file. In particular the Microsoft Excel format, 'xls' is very common. A spreadsheet file can be imported in R using different strategies. A typical approach is to exploit the *RODBC* library (Ripley 2012).

```
#load the RODBC package
library(RODBC)

#open a connection to the spreadsheet file
#using the odbcConnectExcel function
conXL <- odbcConnectExcel("example.xls")

#the sqlTables allows to visualize the database
#structure the field TABLE_NAME, in particular,
#denotes the tables that can be imported from the
#connection object
sqlTables(conXL)

#the function sqlFetch allows to import a data table
#on a worksheet the second argument denotes
#the name of the worksheet to import from a connection
#object (first argument)
dataExample <- sqlFetch(conXL, "sheet1")
```

The function *sqlQuery* allows to exploit the SQL language to query the table: this is useful to partially import the dataset, that is to filter the rows and/or the columns of the table.

```
#the second argument of the sqlQuery function is a
#SQL command
dataExample <- sqlQuery(conXL,
                  "SELECT * FROM [sheet1$sheet1]")
```

```
#in the example the query reads all the data cointaned
#in the worksheet it is possible to query the
#table according to the rows and/or columns
#characteristics using a well specified SQL query
#as second argument
```

There are several other packages on the CRAN repositories proposed to import spreadsheet files. Among the others, the *XLConnect* package (Mirai Solutions GmbH 2012) is interesting as it is completely cross-platform and it does not require an installation of Microsoft Excel, or any special drivers.

```
#load the XLConnect package
library(XLConnect)

#the name of the file to read is the first argument of
#the function readWorksheetFromFile, while the second
#argument denotes the name of the worksheet to read
#the argument header allows to specify that the name of
#the variables are on the first row of the worksheet
dataExample <- readWorksheetFromFile("example.xls",
                                     sheet="sheet1",
                                     header=TRUE)
```

A.1.3 Files from other statistical packages

The *foreign* package (R Core Team 2012a) contains several functions for reading and writing data stored by several statistical packages as well as for reading and writing dBase files.

```
#load the foreign package
library(foreign)

#to import a data set in Stata format
dataExample <- read.dta("example.dta")
#to import a data set in SPSS format
dataExample <- read.spss("example.sav")

#NOTE:
#the function read.ssd and read.xport allows to
#obtain a data frame from a SAS permanent dataset
#the function read.mpt allows to read a Minitable file
#the function read.systat allows to read a Systat file
#the function read.dbf allows to read a DBF file
```

A.2 Exploring data

In this section the main commands for exploring data are shown. In many of the R statements we refer to the dependent variable *y*, in particular for the graphical tools: the study of the distribution of the dependent variable is a key to understand the real added value of applying quantile regression(QR). Nevertheless, what is presented can also be applied to the study of any variable in the dataset.

A.2.1 Graphical tools

While for some graphical representations we refer to the R base graphical system, for most graphs we have exploited both the *ggplot2* (Wickham 2009) and *lattice* (Sarkar 2008) graphical systems. Such systems offer more immediate functions for obtaining complex graphs.

```
#load the ggplot2 package
library(ggplot2)

#load the lattice package
library(lattice)
```

A.2.1.1 Histogram of the dependent variable

```
#histogram using the R traditional graphics system
hist(dataExample$y)

#histogram using the ggplot2 graphics system
ggplot(data=dataExample, aes(y)) + geom_histogram()
#histogram using the lattice graphics system
histogram(~ y, data=dataExample)
```

A.2.1.2 Conditional histograms of the dependent variable

```
#conditional histograms using the ggplot2 graphics
#system with respect to a single categorical variable
ggplot(data=dataExample, aes(y)) +
        geom_histogram() +
        facet_grid(z1 ~ .)

#conditional histograms using the ggplot2 graphics
#system with respect to two categorical variables
ggplot(data=dataExample, aes(y)) +
        geom_histogram() +
        facet_grid(z1 ~ z2)
```

```
#conditional histograms using the lattice graphics
#system with respect to a single categorical variable
histogram(~ y | z1, data=dataExample)

#conditional histograms using the lattice graphics
#system with respect to two categorical variables
histogram(~ y | z1 + z2, data=dataExample)
```

A.2.1.3 Boxplot of the dependent variable

```
#boxplot using the R traditional graphics system
boxplot(dataExample$y)

#boxplot using the ggplot2 graphics system
ggplot(data=dataExample,
       aes(factor(0), y)) +
       geom_boxplot() +
       xlab("") +
       theme(axis.text.x = element_blank())

#boxplot using the lattice graphics system
bwplot(~ y, data=dataExample)
```

A.2.1.4 Conditional boxplots of the dependent variable

```
#conditional boxplots using the R traditional graphics
#system
with(dataExample, boxplot(y ~ z1))

#conditional boxplots using the ggplot2 graphics system
#with respect to a single categorical variable
ggplot(data=dataExample,
       aes(z1, y)) +
       geom_boxplot()
#using different panels
ggplot(data=dataExample,
       aes(factor(0), y)) +
       geom_boxplot() +
       facet_grid(z1 ~ .) +
       xlab("") +
       theme(axis.text.x = element_blank())
```

```
#conditional boxplots using the ggplot2 graphics system
#with respect to two categorical variables using
#different panels
ggplot(data=dataExample,
        aes(factor(0), y)) +
        geom_boxplot() +
        facet_grid(z1 ~ z2) +
        xlab("") +
        theme(axis.text.x = element_blank())

#conditional boxplots using the lattice graphics system
#with respect to a single categorical variable
bwplot(~ y | z1, data=dataExample)

#conditional boxplots using the lattice graphics system
#with respect to two categorical variables
bwplot(~ y | z1 + z2, data=dataExample)
```

A.2.1.5 Quantile plot of the dependent variable

```
#normal quantile plot using the R traditional graphics
#system
qqnorm(y)
#to superimpose the quantile-quantile line on the
#quantile plot
qqline(y, col = 2)

#normal quantile plot using the ggplot graphics system
ggplot(data=dataExample,
        aes(sample = y)) +  stat_qq()

#normal quantile plot using the ggplot graphics system
#superimposing the quantile-quantile line
ggplot(data=dataExample,
        aes(sample = y)) +  stat_qq()  +
        geom_abline(intercept=mean(dataExample$Price),
                    slope=sd(dataExample$Price))

#normal quantile plot using the lattice graphics
#system
qqmath(~ y, data=dataExample, distribution="qnorm")

#to superimpose the quantile-quantile line
```

```
qqmath(~ y, data=dataExample, distribution="qnorm",
        panel = function(x, ...) {
                panel.qqmathline(x, ...)
                panel.qqmath(x, ...)
        })
```

A.2.1.6 Conditional quantile plots of the dependent variable

```
#conditional normal quantile plot using the ggplot
#graphics system
ggplot(data=dataExample,
        aes(sample = y)) +  stat_qq() +
        facet_grid(z1 ~ .)

#conditional normal quantile plot using the lattice
#graphics system
qqmath(~ y | z1, data=dataExample,
       distribution="qnorm")
```

A.2.1.7 Scatter plot

```
#using the R traditional graphics system
with(dataExample, plot(x, y))

#using the ggplot2 graphics system
ggplot(data=dataExample, aes(x, y)) + geom_point()

#using the lattice graphics system
xyplot(data=dataExample, y ~ x1)
```

A.2.1.8 Conditional scatter plots

```
#using the ggplot2 graphics system
#using colours for the different levels of a
#categorical variable
ggplot(data=dataExample, aes(x, y)) +
        geom_point(aes(colour = z1))
#using different panels for the levels of a single
#categorical variable
ggplot(data=dataExample, aes(x, y)) +
        geom_point() +
        facet_grid(z1 ~ .)
```

```
#using different panels for the levels of two
#categorical variables
ggplot(data=dataExample, aes(x, y)) +
        geom_point() +
        facet_grid(z1 ~ z2)

#using the lattice graphics system
#using colours for the different levels of a
#categorical variable
xyplot(data=dataExample, y ~ x1, groups = z1)
#using different panels for the levels of a single
#categorical variable
xyplot(data=dataExample, y ~ x1 | z1)
#using different panels for the levels of two
#categorical variables
xyplot(data=dataExample, y ~ x1 | z1 + z2)
```

A.2.2 Summary statistics

In this section some commands for obtaining the main summary statistics are reported. All the functions shown in the following are available in the R base system, except for the *skewness* and *kurtosis*, which are available in the *moments* package (Komsta and Novomestky 2012).

A.2.2.1 Summary statistics of the dependent variable

```
#main descriptive statistics for all the variables
#in the dataset
summary(dataExample)

#descriptive statistics for a single variable
summary(dataExample$y)

#sample mean, sample variance and standard deviation
#of a numerical variable
mean(dataExample$y)
var(dataExample$y)
sd(dataExample$y)

#5-number summary of a numerical variable
quantile(dataExample$y)

#deciles of a numerical variable
```

```
quantile(dataExample$y, probs=seq(0, 1, 0.1))

#percentiles of a numerical variable
quantile(dataExample$y, probs=seq(0, 1, 0.01))

#skewness and kurtosis of a numerical variable
library(moments)
skewness(dataExample$y)
kurtosis(dataExample$y)
```

A.2.2.2 Conditional summary statistics of the dependent variable

```
#main descriptive statistics
with(dataExample,
        tapply(y, z1, summary))

#sample mean, variance and standard deviation
with(dataExample,
        tapply(y, z1, mean))
with(dataExample,
        tapply(y, z1, var))
with(dataExample,
        tapply(y, z1, sd))

#conditional 5-number summary of a numerical variable
with(dataExample,
        tapply(y, z1, quantile))

#skewness and kurtosis
with(dataExample,
        tapply(y, z1, skewness))
with(dataExample,
        tapply(y, z1, kurtosis))

#deciles of a numerical variable
with(dataExample,
        tapply(y, z1, quantile, probs=seq(0, 1, 0.1)))

#percentiles of a numerical variable
with(dataExample,
        tapply(y, z1, quantile, probs=seq(0, 1, 0.01)))
```

```
#NOTE
#the conditional statistics can be obtained with respect
#to two or more categorical variables using a list of
#variables as second argument example for the mean:
with(dataExample,
        tapply(y, list(z1, z2), mean))
```

A.2.2.3 Contingency tables

```
#frequency table of a categorical variable
#using the count
table(dataExample$z1)
#using the proportions
prop.table(table(dataExample$z1))
#contingency table of two categorical variables
#using the count
with(dataExample,
        table(z1, z2))
#using the proportions
with(dataExample,
        prop.table(table(z1, z2)))
```

A.3 Modeling data

A.3.1 Ordinary least squares regression analysis

A.3.1.1 Parameter estimates

```
#OLS regression
olsModel <- lm(y ~ x, data=dataExample)
#printing the OLS results
summary(olsModel)

#OLS fitted values
fitted(olsModel)

#OLS residuals
residuals(olsModel)

#OLS regression introducing a categorical explanatory
#variable
```

```
olsModel2 <- lm(data=dataExample, y ~ x1 + z1)
summary(olsModel2)

#OLS regression: an example of multiple regression
#model
olsModel3 <- lm(data=dataExample, y ~ x1 + x2 + x3 +
                x4 + z1)
summary(olsModel3)
```

A.3.1.2 Main graphical representations

```
#OLS regression line using the R traditional graphics
#system
plot(data=dataExample, y ~ x)
abline(lm(data=dataExample, y ~ x))
#main OLS plots:
# 1.fitted values vs residuals
# 2.normal quantile plot of the standardized residuals
# 3.fitted values vs square root of the standardized
#   residuals
# 4.leverage vs standardized residuals
plot(olsModel)
```

A.3.2 Quantile regression analysis

The functions for carrying out a complete QR analysis are available in the *quantreg* package (Koenker 2011). The interested reader is strongly encouraged to consult the manual and the vignette of the package for more examples. The function *melt* hosted in the *reshape2* package (Wickham 2007) has been used to prepare the data for some graphical representations.

```
#loading the quantreg package
library(quantreg)

#load the reshape2 package, to change the data table
#format
library(reshape2)
```

A.3.2.1 Parameter estimates

```
#QR regression using a single conditional quantile
#(the conditional median)
```

```
qrModel <- rq(data=dataExample, y ~ x, tau=0.5)
#print the QR results
summary(qrModel)
#to obtain the variance/covariance matrix
summary(qrModel, covariance=TRUE)$cov

#QR regression using a set of conditional quantiles
qrModel <- rq(data=dataExample, y ~ x,
                tau= c(0.1, 0.25,0.5,0.75, 0.9))

#QR regression estimating the whole quantile process
qrModel <- rq(data=dataExample, y ~ x, tau=-1)
#printing the QR results
qrModel <- rq(data=dataExample, y ~ x,
                tau=c(0.1,0.25,0.5,0.75,0.9))
summary(qrModel)

#QR fitted values
fitted(qrModel)

#QR residuals
residuals(qrModel)
```

The default method used for estimating conditional quantiles exploits a modified version of the simplex algorithm (see Chapter 2 for details and more references). It is possible to specify different estimation algorithms by setting the *method* option.

```
#QR regression exploiting the simplex method
qrModel <- rq(data=dataExample, y ~ x, tau=0.5,
              method="br")

#QR regression exploiting Frisch-Newton approach
qrModel <- rq(data=dataExample, y ~ x, tau=0.5,
              method="fn")

#QR regression exploiting Frisch-Newton approach
#after preprocessing
qrModel <- rq(data=dataExample, y ~ x, tau=0.5,
              method="pfn")
#NOTE:
#the option "fnc" allows to specify linear inequality
```

```
#    constraints on the fitted coefficient
#the option "lasso" implements the lasso penalty
#the option "scad" implements the Fan and Li's smoothly
#    clipped absolute deviation penalty
```

A.3.2.2 Main graphical representations

```
#to extract the QR residuals
qrResid <- residuals(qrModel)

#the function melt allows to change the data table
#format (from a wide format to a long format)
qrmResid <- melt(qrResid, variable.name="residuals")
#plot of the residuals using panels for the different
#conditional quantile models stored in the QR object
#using the ggplot2 graphics system
qplot(data=qrmResid, x=Var1, y=residuals,
      geom="point") + facet_grid(~ Var2)

#QR coefficients plot using the R graphics system
plot(qrModel)

#QR lines using the R graphics system
plot(data=dataExample, y ~ x)
taus <- c(0.1, 0.25, 0.5, 0.75, 0.9)
for(i in 1:length(taus)){
    abline(data=dataExample, y ~ x, tau=taus[i]),
           col="gray")
}

#QR lines using the ggplot2 graphical system
#using 5 quantiles
ggplot(data=dataExample, aes(x1, y)) + geom_point() +
       stat_quantile(quantiles = c(0.1, 0.25, 0.5,
                     0.75, 0.9))
#using a sequence of quantiles represented using
#different gradations of colour
ggplot(data=dataExample, aes(x1, y)) + geom_point() +
       stat_quantile(aes(colour = ..quantile..),
                     quantiles = seq(0.05, 0.95,
                                     by=0.05))
```

A.3.2.3 QR coefficients introducing a categorical explanatory variable

```
#QR estimates introducing a categorical variable in
#the model
qrModel2 <- rq(data=dataExample, y ~ x1 + z1)
summary(qrModel2)
```

A.3.2.4 QR coefficients for a multiple regression model

```
#QR estimates for a multiple regression model
qrModel2 <- rq(data=dataExample, y ~ x1 + x2 + x3 +
               x4 + z1)
```

A.3.2.5 Confidence intervals

```
#set of quantiles of interest
taus <- c(0.25,0.5,0.75)
#QR estimation
fit_rq <-rq(formula = y ~ x, taus)
#confidence intervals computed by the rank inversion
#method
summary(fit_rq)
#confidence intervals iid errors
summary(fit_rq, se="iid")
#confidence intervals nid errors
summary(fit_rq, se="nid")
#confidence intervals computed by the bootstrap
#xy-pair method
summary(fit_rq, se="boot", bsmethod = "xy")
#confidence intervals computed by the bootstrap
#Parzen method
summary(fit_rq, se="boot", bsmethod = "pwy")
#confidence intervals computed by the bootstrap
#MC method
summary(fit_rq, se="boot", bsmethod = "mcmb")
#to save the bootstrap coefficients
bootCoefficients <- boot.rq(y, x, tau = 0.5, R = 500,
                            bsmethod = "xy")
```

A.3.2.6 Hypothesis tests

The function *anova.rq* allows different hypothesis tests to be conducted. In particular, if all the fitted objects have the same specified quantile but different regressors,

the intent is to test that smaller models are adequate relative to the largest specified model. It is possible to use the generic function *anova*: when a *rq* object is passed the proper function is used.

```
#joint test of equality of slopes
fit_1var <- rq(y ~ x1, tau = 0.25)
fit_2vars <- rq(y ~ x1 + x2, tau = 0.25)
fit_3vars <- rq(y ~ x1 + x2 + x3, tau = 0.25)
#joint tests of equality on all slope parameters
#using a Wald test (default)
anova(fit_1var, fit_2var, fit_3var)
anova(fit_1var, fit_2var, fit_3var, test ="Wald")
#joint tests of equality on all slope parameters
#using a rank inversion method
anova(fit_1var, fit_2var, fit_3var, test ="rank")
#using the method based on the analysis of the
#weighted residuals: the test is based on the difference
#in the QR objective functions at the restricted and
#unrestricted models with a reference distribution
#computed by simulation
anova(fit_1var, fit_2var, fit_3var, test ="anowar")

#separate tests on each of the slope parameters
#using a Wald test (default)
anova(fit_1var, fit_2var, fit_3var)
anova(fit_1var, fit_2var, fit_3var, joint=FALSE)
anova(fit_1var, fit_2var, fit_3var, joint=FALSE,
      test="Wald")
#separate tests on each of the slope parameters
#using a rank inversion method
#it is possible to specify a form for the score
#function: the Wilcoxon score is used as default
anova(fit_1var, fit_2var, fit_3var, joint=FALSE,
      test="rank")
anova(fit_1var, fit_2var, fit_3var, joint=FALSE,
      test="rank", score="Wilcoxon")
#median (sign) scores
anova(fit_1var, fit_2var, fit_3var, joint=FALSE,
      test="rank", score="sign")
#normal scores
anova(fit_1var, fit_2var, fit_3var, joint=FALSE,
      test="rank", score="normal")
#generalization of median scores to an arbitrary
#quantile: the quantile is assumed to be the one
#associated with the fitting of the specified objects
anova(fit_1var, fit_2var, fit_3var, joint=FALSE,
      test="rank", score="tau")
```

In the case where the fitted object passed to the *anova.rq* function refers to the same regressors but the specified quantiles are different, the intent is to verify that the linear predictor of the fits are all the same for the several quantiles.

```
#separate tests on each of the slope parameters
#using a Wald test (default)
anova(fit_25, fit_50, fit_75)
anova(fit_25, fit_50, fit_75, joint=TRUE)
anova(fit_25, fit_50, fit_75, joint=TRUE, test="Wald")
#separate tests on each of the slope parameters
#using a rank inversion method
anova(fit_25, fit_50, fit_75, joint=TRUE, test="rank")
#the same option shown above can be used to set
#the score function
```

A.4 Exporting figures and tables

A.4.1 Exporting figures

```
#exporting figures created in the R base graphical
#system using the pdf device
#(it is possible to set the desired size of the image)
pdf("graph1.pdf")
#histogram using the R traditional graphics system
hist(dataExample$y)
dev.off()
#NOTE: main devices available
# postscript(): Adobe PostScript file
# pictex(): LATEXPicTeX file
# xfig(): XFIG file
# bitmap(): GhostScript conversion to file
# png(): PNG bitmap file
# jpeg(): JPEG bitmap file
#the same devices can be used for saving ggplot2 and
#lattice graphics
#the ggplot2 package contains a convenient function
#for saving a plot for a default size using the size of
#the current graphics device. It is also possible to set
#the desired size of the image
#the function guesses the type of graphics device from
#the file extension
```

```
#ggsave by default saves the last plot displayed
#example: histogram using the ggplot2 graphics system
ggplot(data=dataExample, aes(y)) + geom_histogram()
ggsave(file="graph1.pdf")

#if the plot is assigned to an R object
ggp <- ggplot(data=dataExample, aes(y)) +
              geom_histogram()
#it is possible to specify the object as first
#argument of the function
ggsave(ggp, file="graph1.pdf")
```

A.4.2 Exporting tables

```
#the function xtable is available in the xtable package
#(Dahl, 2012)
#load the xtable package
library(xtable)
#to export tables in latex format
xtable(summary(lm(y ~ x1, data=dataExample)),
       type="latex")
#to export tables in html format
xtable(summary(lm(y ~ x1, data=dataExample)),
       type="html")
#NOTE:
#the package quantreg has a latex.table function for
#exporting results to LaTeX
#for exporting results to different formats (LaTeX, doc,
#pdf, html, xml, etc.) see the Sweave function,
#available in the R base system and the knitr  package
#(Xie, 2012), among the others
```

References

Cohen H and Cohen H 2008 *Statistics and Data with R. John* Wiley & Sons, Ltd.

Dahl DB 2012 *xtable: Export Tables to LaTeX or HTML*. R package version 1.7. http://CRAN.R-project.org/package=xtable.

Dalgaard P 2008 *Introductory Statistics with R*, 2nd Edition. Springer.

Koenker R 2011 *quantreg: Quantile Regression*. R package version 4.76. http://CRAN.R-project.org/package=quantreg.

Komsta L and Novomestky F 2012 *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.13. http://CRAN.R-project.org/package=moments.

Mirai Solutions GmbH 2011 *XLConnect: Excel Connector for R*. R package version 0.2-0. http://CRAN.R-project.org/package=XLConnect.

R Core Team 2012a *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...* R package version 0.8-50. http://CRAN.R-project.org/package=foreign.

R Core Team 2012b *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

R Core Team 2012c *R Data Import/Export*. R Foundation for Statistical Computing. http://www.R-project.org/.

Ripley B (and from 1999 to October 2002 Lapsley M) 2012 *RODBC: ODBC Database Access*. R package version 1.3-6. http://CRAN.R-project.org/package=RODBC.

Sarkar D 2008 *Lattice: Multivariate Data Visualization with R*. Springer.

Teetor P 2011 *R Cookbook*. O'Reilly.

Venables WN and Ripley BD 2002 *Modern Applied Statistics with S*, 4th Edition. Springer.

Wickham H 2007 Reshaping data with the reshape package. *Journal of Statistical Software* **21**(12), 1–20.

Wickham H 2009 *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Xie Y 2012 *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 0.8. http://CRAN.R-project.org/package=knitr.

Appendix B

Quantile regression and surroundings using SAS

Introduction

This appendix is devoted to the presentation of the main commands available in SAS for carrying out a complete data analysis, that is for loading, exploring and modeling data. As in Appendix A we assume that the data file is stored in the current working directory in a file named *example.{txt, csv, xls, dta, sav}* where the file extensions are associated with the following:

txt tab delimited text file;
csv comma separated values file;
xls spreadsheet (Microsoft Excel) file;
dta Stata file;
sav Spss data file.

Moreover, we assume that the file contains a dependent numerical variable (y), a set of numerical explanatory variables ($x1, x2, x3, x4$), and a set of categorical explanatory variables ($z1, z2$).

Throughout this appendix, SAS commands are shown in bold font and comments using regular font. The main commands and their details are fully presented in the official SAS documentation (SAS Institute Inc. 2010a). A further interesting reading is the evergreen book of Delwiche and Slaughter (2012), continuously updated to take into account the new features implemented in the subsequent versions of the software. It is worth to mention the two introductory books on SAS by Cody

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

(2007, 2011). Finally, the book by Kleinman and Horton (2009) offers a guide to data management, statistical analysis and graphics through a continuous comparison of R and SAS.

B.1 Loading data

B.1.1 Text data

The functions *INFILE* and *IMPORT* allow to import a text file. Each function is able to read both tab delimited files and comma separated values files.

```
/* to import data in tab separated text format */
DATA dataExample;
      INFILE 'example.txt' DLM='09'X DSD;
      INPUT y x1 x2 x3 x4 z1 z2;
RUN;
PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;

/*
a different way of reading delimited files exploits the
IMPORT procedure (if the first line does not contain the
column headers set GETNAMES = NO)
SAS determines the file type by the extension (use the
DELIMITER option to specify a different column delimiter)
*/
PROC IMPORT DATAFILE = 'example.txt' OUT = dataExample REPLACE;
      GETNAMES = YES;
RUN;
PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;
```

In the case of comma separated values files, the only option to change for the *INFILE* procedure is the *DLM* argument.

The *IMPORT* procedure automatically determines the file type through the file extension, although it is possible to specify a custom delimiter not associated with the file extension, by specifying the *DELIMITER* option.

```
/* to import csv data (comma separated values) */
DATA dataExample;
      INFILE 'example.csv' DLM=',' DSD;
```

```
      INPUT y x1 x2 x3 x4 z1 z2;
RUN;

PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;

/*
a different way of reading delimited files exploits the
IMPORT procedure (if the first line does not contain the
column headers, set GETNAMES = NO)
SAS determines the file type by the extension (use the
DELIMITER option to specify a different column delimiter)
*/
PROC IMPORT DATAFILE = 'example.csv' OUT = dataExample
      REPLACE;
      GETNAMES = YES;
RUN;
PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;
```

B.1.2 Spreadsheet data

The SAS/ACCESS interface allows to read spreadsheet data, exploiting again the *IMPORT* command.

```
/*
it is possible to read spreadsheet files if the SAS/ACCESS
interface is installed
*/
PROC IMPORT DATAFILE = 'example.xls' OUT = dataExample
      DBMS = XLS SHEET = "sheet1" REPLACE;
RUN;
PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;
```

B.1.3 Files from other statistical packages

The *IMPORT* procedure allows to read both *Stata* (.dta) and *SPSS* (.sav) files if the SAS/ACCESS interface to PC file formats is installed.

```
/*
it is possible to read SPSS files if the SAS/ACCESS
interface is installed
*/
PROC IMPORT DATAFILE = 'example.sav' OUT = dataExample
      DBMS = SAV REPLACE;
RUN;
PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;

/*
the same procedure allows to read also a Stata data file
by specifying the proper DMBS option
*/
PROC IMPORT DATAFILE = 'example.dta' OUT = dataExample
      DBMS = DTA REPLACE;
RUN;
PROC PRINT DATA = dataExample;
      TITLE 'Dataset of example';
RUN;
```

B.2 Exploring data

B.2.1 Graphical tools

Graphical representations can be obtained both exploiting the *SGPLOT* procedure and the *ODS graphics* system. The latter is an extension of the *Output Delivery System* and requires the *SAS/GRAPH* software to be installed. For most of the graphs referred to in this appendix, the commands in the two graphical systems are provided.

B.2.1.1 Histogram of the dependent variable

```
/*
histogram exploiting the SGPLOT procedure
*/
PROC SGPLOT DATA=dataExample;
      HISTOGRAM y;
      TITLE 'Histogram of y';
RUN;

/*
histogram exploiting the ODS GRAPHICS system
*/
```

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = dataExample;
      VAR y;
      HISTOGRAM y;
      TITLE;
RUN;

/*
ODS Graphics remains in effect for all procedures that follow
although it is not necessary to turn ODS graphics off, in
order to save resources it is possible to turn it off
through the dual command
*/
ODS GRAPHICS OFF;
```

Finally, it is possible to obtain a histogram also exploiting the *UNIVARIATE* procedure by specifying the proper option.

```
/*
to obtain the histogram from the UNIVARIATE procedure
*/
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = dataExample;
      VAR y;
      HISTOGRAM y/NORMAL;
      TITLE;
RUN;
```

B.2.1.2 Conditional histograms of the dependent variable

```
/*
histogram of a numerical variable conditional on the
different levels of a categorical variable, exploiting the
SGPLOT procedure
*/
PROC SGPLOT DATA=dataExample;
      HISTOGRAM y / CATEGORY z1;
      TITLE 'Histogram of y conditioned on z1';
RUN;

/*
and exploiting the ODS Graphics system through the UNIVARIATE
procedure
*/
```

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = dataExample;
      VAR y;
      HISTOGRAM y;
      BY z1;
      TITLE 'Histogram of y conditioned on z1';
RUN;
```

B.2.1.3 Boxplot of the dependent variable

```
/*
boxplot of a numerical variable
*/
PROC SGPLOT DATA=dataExample;
      HBOX y;
      TITLE 'Boxplot of y';
RUN;
```

B.2.1.4 Conditional boxplots of the dependent variable

```
/*
boxplot of a numerical variable conditioned on the different
levels of a categorical variable through the SGPLOT procedure
*/
PROC SGPLOT DATA=dataExample;
      VBOX y / CATEGORY z1;
      TITLE 'Boxplot of y' conditioned on z1;
RUN;
```

B.2.1.5 Quantile plot of the dependent variable

```
/*
quantile plot of a numerical variable
*/
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = dataExample;
      VAR y;
      PROBPLOT y;
      BY z1;
      TITLE 'Normal probability plot';
RUN;
```

```
/*
the plot uses the normal distribution by default
it is possible to set a different reference distribution
specifying it with a plot option
the available options are: BETA, EXPONENTIAL, GAMMA,
LOGNORMAL, NORMAL and WEIBULL
*/
```

B.2.1.6 Conditional quantile plots of the dependent variable

```
/*
quantile plot of a numerical variable conditioned on the
different levels of a categorical variable through the
UNIVARIATE procedure setting the BY option
*/
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = dataExample;
      VAR y;
      PROBPLOT y;
      BY z1;
      TITLE 'Normal probability plots';
RUN;
```

B.2.1.7 Scatter plot

```
PROC SGPLOT DATA=dataExample;
      SCATTER x=x1 y=y;
RUN;
```

B.2.1.8 Conditional scatter plots

```
/*
the GROUP option allows to specify a third variable to be
used for grouping the data
*/
PROC SGPLOT DATA=dataExample;
      SCATTER x=x1 y=y / GROUP=z1;
RUN;
```

B.2.2 Summary statistics

Summary statistics for a single variable or for a set of variables are available through the *MEANS*, the *UNIVARIATE* or the *TABULATE* procedures.

B.2.2.1 Summary statistics of the dependent variable

```
/*
by default, the procedure MEANS prints the number of
non-missing values, the mean, the standard deviation, the
minimum and the maximum values for each specified variable
*/
PROC MEANS DATA=dataExample;
      VAR y x1 x2 x3 x4;
      TITLE 'Summary statistics for all the variables';
RUN;

PROC MEANS DATA=dataExample;
      VAR y;
      TITLE 'Summary statistics for a single variable';
RUN;

/*
it is possible to specify the statistics to print as options
to the procedure: minimum, average, median, maximum,
standard deviation and kurtosis
*/
PROC MEANS DATA=dataExample
      MIN MEAN MEDIAN MAX STDEV SKEWNESS KURTOSIS;
      VAR y x1 x2 x3 x4;
      TITLE 'Main summary statistics';
RUN;

/*
the three quartiles
*/
PROC MEANS DATA=dataExample
      Q1 MEDIAN Q3;
      VAR y;
      TITLE 'The three quartiles for the dependent variable';
RUN;

/*
the nine deciles
*/
PROC MEANS DATA=dataExample
      P10 P20 P30 P40 P50 P60 P70 P80 P90;
      VAR y;
```

```
      TITLE 'The deciles for the dependent variable';
RUN;
```

```
/*
the procedure MEANS use 0.05 as default confidence limit
it is possible to set a different value using the ALPHA
option along with the CLM option
*/
PROC MEANS DATA=dataExample
      ALPHA=.10 CLM;
      VAR y;
      TITLE 'Summary statistics using a different
            confidence limit';
RUN;

/*
the UNIVARIATE procedure provides statistics and graphs
for all the numeric variables in a data set. The statistics
include mean, median, mode, standard deviation, skewnees,
kurtosis, range, interquartile range, main quantiles
*/
PROC UNIVARIATE DATA=dataExample;
      TITLE 'A different procedure for summary statistics';
RUN;

/*
using the VAR statement, it is possible to limit the
statistics to a single or more numerical variables
*/
PROC UNIVARIATE DATA=dataExample;
      VAR y;
      TITLE 'A different procedure for summary statistics';
RUN;
```

B.2.2.2 Conditional summary statistics of the dependent variable

```
/*
summary statistics for groups exploiting the MEANS procedure
the CLASS statement allows to obtain summary statistics of a
numerical variable conditioned on the different levels of a
categorical variable
*/
PROC MEANS DATA=dataExample;
      VAR y;
      CLASS z1;
```

```
      TITLE 'Summary statistics for groups';
RUN;

/*
an alternative exploits the TABULATE procedure again
specifying a CLASS statement
*/
PROC TABULATE DATA=dataExample;
      CLASS z1;
      TABLE y ALL, MIN MEDIAN MEAN MAX STDDEV *y*(z1 ALL);
      TITLE 'Summary statistics for groups';
RUN;
```

B.2.2.3 Contingency tables

```
/*
the FREQ procedure allows to summarize categorical variables
through frequency univariate table
*/
PROC FREQ DATA=dataExample
      TABLES z1;
      TITLE 'Frequency table for the categorical variable';
RUN;

/*
contingency table
statistical options allow to test categorical data using
different indexes
*/
PROC FREQ DATA=dataExample
      TABLES z1 * z2;
      TITLE 'Contingency table for two categorical variables';
RUN;
```

B.3 Modeling data

B.3.1 Ordinary least squares regression analysis

B.3.1.1 Parameter estimates

```
/*
the REG procedure fits linear regression models by
least-squares
the MODEL dependent = independent statement allows
to specify the analysis model
*/
```

```
PROC REG DATA=dataExample;
      MODEL y=x1;
      TITLE 'Results of OLS Regression Analysis';
RUN;
```

To introduce a categorical variable in the model, it is possible to code the levels of the variable through dummy variables or to exploit the *GLM* procedure. The *GLM* procedure generates dummy variables for a categorical variable on-the-fly without the need to code the variable manually.

```
/*
simple regression model using a nominal regressor
the CLASS statement specifies that the variable z1 is a
categorical variable
the option ORDER = FREQ orders the levels of the class
variable according to descending frequency count:
levels with the most observations come first in the order
the solution option allows to obtain the parameter estimates
the SS3 option specifies that Type III sum of squares is
used for hypothesis test
*/
PROC GLM DATA = dataExample ORDER=FREQ;
  CLASS z1;
  MODEL y = z1 / SOLUTION SS3;
RUN;
QUIT;
```

```
/*
multiple regression model
*/
PROC REG DATA=dataExample;
      MODEL y=x1 x2 x3 x4;
      TITLE 'Results of OLS Regression Analysis';
RUN;

/*
multiple regression model
the STB option of the MODEL statement requests a table of
the standardized values (beta coefficients)
*/
PROC REG DATA=dataExample;
      MODEL y=x1 x2 x3 x4 / STB;
      TITLE 'Results of OLS Regression Analysis';
RUN;
```

```
/*
multiple regression model test on a single coefficient
(the test1 is a label to identify the output of the
 test command)
*/
PROC REG DATA=dataExample;
      MODEL y=x1 x2 x3 x4;
      TITLE 'Results of OLS Regression Analysis';
      test1: x1 = 0;
RUN;

/*
the same results is obtained using:
      test1: x1;
since SAS defaults to comparing the term(s) listed to 0.
*/

/*
multiple regression model test on multiple coefficients
(the test2 is a label to identify the output of the
 test command)
*/
PROC REG DATA=dataExample;
      MODEL y=x1 x2 x3 x4;
      TITLE 'Results of OLS Regression Analysis';
      test2: x1 x2 x3 x4;
RUN;
```

B.3.1.2 Main graphical representations

```
/*
scatter plot with regression line, confidence and prediction
bands superimposed
*/

PROC REG DATA=dataExample
      PLOTS(ONLY) = (FITPLOT);
      MODEL y=x1;
      TITLE 'Scatter, OLS line, confidence and prediction
             bands';
RUN;
```

```
/*
residuals vs independent variable
*/
```

```
PROC REG DATA=dataExample
      PLOTS(ONLY) = (RESIDUALS);
      MODEL y=x1;
      TITLE 'Residual plots for OLS Regression Analysis';
RUN;
```

```
/*
eight different diagnostic plots arranged in a panel plot
*/
PROC REG DATA=dataExample
      PLOTS(ONLY) = (DIAGNOSTICS);
      MODEL y=x1;
      TITLE 'Eight diagnostic plots for OLS Regression
            Analysis';
RUN;
```

```
/*
It is possible to obtain a single plot using the proper
options (see the capitalized values in the following list)
COOKSD: observation number vs Cook's D statistic
OBSERVEDBYPREDICTED: predicted values vs dependent values
QQPLOT: normal quantile plot of residuals
RESIDUALBYPREDICTED: predicted values vs residuals
RESIDUALHISTOGRAM: histogram of residuals
RFPLOT: residual fit plot
RSTUDENTBYLEVERAGE: leverage vs studentized residuals
RSTUDENTBYPREDICTED: predicted values vs studentized residuals
*/
```

```
/*
it is possible to specificy a list of desired plots to
the PLOTS option
*/
PROC REG DATA=dataExample
      PLOTS(ONLY) = (RESIDUALS QQPLOT RESIDUALHISTOGRAM);
      MODEL y=x1;
      TITLE 'Residual plots for OLS Regression Analysis';
RUN;
```

B.3.2 Quantile regression analysis

Quantile regression in SAS is carried out by the *QUANTREG* procedure (SAS Institute Inc. 2010b).

```
/*
quantile regression at a specified conditional quantile
*/
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE=0.1;
      TITLE 'Results of QR Regression Analysis';
RUN;

/*
quantile regression in correspondence of a set of
conditional quantiles
*/
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE=0.1 0.25 0.5 0.75 0.9;
      TITLE 'Results of QR Regression Analysis';
RUN;

/*
the default estimation algorithm depends on the size of the
problem (number of observations and number of covariates)
it is possible to specify the estimation algorithm by
setting the ALGORITHM option to:
- SIMPLEX (simplex method)
- INTERIOR (interior point method)
- SMOOTH (smoothing algorithm)
*/

/*
the QUANTILE=process option allows to estimate the
whole quantile process
the PLOT=quantplot option requests a graphical representation
of the quantile regression estimates
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE=process
      PLOT=quantplot;
RUN;
ODS graphics off;
```

The covariance and the correlation matrices for the estimated regression coefficients are available through *COVB* and *CORRB*, respectively. The method used to

compute the matrices depends on the method used to specify the confidence intervals (see Section B.3.2.1 for the available options).

```
/*
the COVB option requests covariance matrices for the
estimated regression coefficients
- the bootstrap covariance is computed when the resampling
method is used to compute the confidence intervals
- the asymptotic covariance based on an estimator of the
sparsity function is computed when the sparsity method is
used to compute the confidence intervals
- the rank method for confidence intervals does not provide a
covariance estimate
*/
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE = 0.25 0.5 0.75 COVB;
RUN;

/*
the CORRB option requests correlation matrices for the
estimated regression coefficients
- the bootstrap correlation is computed when the resampling
method is used to compute the confidence intervals
- the asymptotic correlation based on an estimator of the
sparsity function is computed when the sparsity method is
used to compute the confidence intervals
- the rank method for confidence intervals does not provide a
correlation estimate
*/
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE = 0.25 0.5 0.75 CORRB;
RUN;
```

B.3.2.1 Confidence intervals and hypothesis tests

```
/*
the RANK option requests to compute the confidence
intervals by exploiting the inversion of rank-score tests
the option CI along with the statement ALPHA allows to
compute the confidence intervals at the level 0.9
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample alpha=0.1 CI = RANK;
      MODEL y = x1 / QUANTILE = 0.1 0.25 0.5 0.75 0.9
PLOT = quantplot;
```

```
RUN;
ODS graphics off;

/*
the SPARSITY option exploits the sparsity function
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample alpha=0.1 CI = SPARSITY;
      MODEL y = x1 / QUANTILE = 0.1 0.25 0.5 0.75 0.9
PLOT = quantplot;
RUN;
ODS graphics off;

/*
two suboption are available for estimating the sparsity
function:
- by specifying IID as suboption, the sparsity function is
estimated assuming that the errors in the linear model are
independent and identically distributed
- by specifying HS (Hall-Sheather method) or BF (Bofinger
method), the sparsity function is estimated assuming
that the conditional quantile function is locally linear and
a bandwith selection method is used
(by default, the Hall-Sheather method is used)
*/

/*
the RESAMPLING option requests the use of the resampling
method to compute confidence intervals
the NREP suboption allows to specify the number of repeats
(by default, NREP=200 - the value of NREP must be greater
than 50)
the SEED option specifies a seed for the resampling method
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample alpha=0.1 CI=RESAMPLING;
      MODEL y = x1 / QUANTILE = 0.1 0.25 0.5 0.75 0.9
      SEED =123
PLOT = quantplot;
RUN;
ODS graphics off;

/*
by specifying the CI=resampling option, QUANTREG also computes
standard errors, t-statistics, and p-values.
*/
```

The confidence intervals can be estimated also when the whole quantile process is computed.

```
/*
the QUANTILE = process option requests an estimate of the
whole quantile process for each regression parameter the
ALPHA = 0.1 option and CI = resampling request to compute 90%
confidence bands for the quantile process using the
resampling method
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample alpha=0.1 CI=resampling;
      MODEL y = x1 / QUANTILE = 0.1 0.25 0.5 0.75 0.9
      SEED =123
PLOT = quantplot;
RUN;
ODS graphics off;
```

To obtain a test for the canonical linear hypothesis concerning the parameters, the *TEST* statement is available. The tested effects can be any set of effects in the *MODEL* statement. The test procedure exploits the Wald test method, the rank inversion method and the likelihood ratio method.

```
/*
by specifying the TEST statement, tests of significance on
the slope parameters are computed
the option WALD in the TEST statement requests
Wald tests
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75 COVB seed=123;
      TEST x1 / WALD;
RUN;
ODS graphics off;

/*
the option LR in the TEST statement requests likelihood
ratio tests
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75 COVB seed=123;
      TEST x1 / LR;
```

```
RUN;
ODS graphics off;

/*
the option RANKSCORE in the TEST statement requests rank
score tests
the following three score functions are available:
- RANKSCORE(NORMAL): normal scores (default)
- RANKSCORE(WILCOXON): Wilcoxon scores
- RANKSCORE(SIGN): sign scores
they are asymptotically optimal for the Gaussian,
logistic, and Laplace location shift models, respectively
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75 COVB seed=123;
      TEST x1 / RANKSCORE;
RUN;
ODS graphics off;

/*
it is possible to specify multiple option to the
TEST statement to obtain different type of tests
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75 COVB seed=123;
      TEST x1 / WALD LR;
RUN;
ODS graphics off;
```

B.3.2.2 Main graphical representations

```
/*
by default the QUANTREG procedure creates the quantile
fit plot
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75;
RUN;
ODS graphics off;

/*
the PLOTS = NONE option requests to suppress all plots
*/
```

```
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75;
      PLOTS = NONE;
RUN;
ODS graphics off;

/*
in the case of a single regression model with a continuous
regressor, the PLOTS = FITPLOT option creates a plot of fitted
conditional quantlles againts the single continuous variable,
using a line for each conditional quantile
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75;
      PLOTS = FITPLOT;
RUN;
ODS graphics off;

/*
the PLOTS = DDPLOT option creates a plot of robust distance
against Mahalanobis distance
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75
      PLOTS = DDPLOT;
RUN;
ODS graphics off;

/*
the PLOTS = HISTOGRAM option creates a histogram for the
standardized residuals based on the quantile regression
estimates, superimposing a normal density curve and a kernel
density curve on the plot
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75
      PLOT = HISTOGRAM;
RUN;
ODS graphics off;

/*
the PLOTS = QQPLOT option creates a normal quantile-quantile
plot for the standardized residuals based on the quantile
regression estimates
```

```
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75
      PLOT = QQPLOT;
RUN;
ODS graphics off;

/*
the PLOTS = RDPLOT option creates a plot of standardized
residuals against robust distance
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75;
      PLOTS = RDPLOT;
RUN;
ODS graphics off;

/*
the PLOTS = ONLY option suppresses the default quantile fit
plot and allows to specify one or more plot to display (in
the example the histogram and the quantile-quantile plot of
the standardized residuals)
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75;
      PLOTS = ONLY (HISTOGRAM QQPLOT);
RUN;
ODS graphics off;

/*
the PLOTS = ALL creates all appropriate plots
*/
ODS graphics on;
PROC QUANTREG DATA=dataExample ALPHA=0.1 CI=resampling;;
      MODEL y=x1 / QUANTILE=0.25 0.5 0.75;
      PLOTS = ALL;
RUN;
ODS graphics off;
```

B.4 Exporting figures and tables

Output management in SAS is managed by the *Output Delivery System (ODS)*. *ODS* determines the different type of output that can be produced, the so-called destinations.

The default destination for the output is the *LISTING* destination, that is the output window or an output file in the case of batch mode. It is possible to change the destination to create proper output in a different format. The available destinations include:

LISTING standard SAS output;
DOCUMENT general output document;
HTML HTML document;
MARKUP markup document (different languages can be specified);
RTF RTF document;
PDF Adobe PDF document;
PS Adobe PostScript document.

The *DOCUMENT* destination can be used as the general format in the case the final format of the report has not yet been fixed. The particular format used for the destinations is determined by templates. The different destinations are associated with default templates but it is possible to specify a different template or to create a custom template.

```
/*
to list the available templates
*/
PROC TEMPLATE;
      LIST STYLES;
RUN;
```

It is possible to select or exclude output objects using the *ODS SELECT* and *ODS EXCLUDE* statements, respectively. In order to determine the name of the objects in the output, the *ODS TRACE ON/OFF* statements are available (SAS Institute Inc. 2010).

The *ODS* statements are global and do not belong to either *DATA* or *PROC* steps: it is advisable to put the opening *ODS* statement just before the step(s) to capture and the closing *ODS* statement following the *RUN* statement.

```
/*
to use an HTML destination
the ODS NOPROCTITLE statement allows to remove procedure
names form output
*/
ODS HTML FILE = 'c:\outputDirectory\outputExample.html';
ODS NOPROCTITLE;
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE=0.1 0.25 0.5 0.75 0.9;
      TITLE 'Results of QR Regression Analysis';
```

```
RUN;

/*
to close the HTML file
*/
ODS HTML CLOSE;

/*
to use a RTF destination
the BODYTITLE statement puts titles and footnotes in the
main part of the RTF document and not in headers or footers
*/
ODS RTF FILE = 'c:\outputDirectory\outputExample.rtf'
BODYTITLE;
ODS NOPROCTITLE;
PROC QUANTREG DATA=dataExample;
      MODEL y=x1 / QUANTILE=0.1 0.25 0.5 0.75 0.9;
      TITLE 'Results of QR Regression Analysis';
RUN;

/*
to close the RTF file
*/
ODS RTF CLOSE;
```

References

Cody R 2007 *Learning SAS by Example: A Programmer's Guide*. SAS Publishing.

Cody R 2011 *SAS Statistics by Example*. SAS Publishing.

Delwiche L and Slaughter S 2012 *The Little SAS Book*. SAS Institute.

Kleinman K and Horton NJ 2009 *SAS and R: Data Management, Statistical Analysis and Graphics*. Chapman & Hall / CRC.

SAS Institute Inc. 2010a *SAS/STAT 9.22 User's Guide*. SAS Institute Inc.

SAS Institute Inc. 2010b *SAS/STAT 9.22 User's Guide. The QUANTREG Procedure*. SAS Institute Inc.

Appendix C

Quantile regression and surroundings using Stata

Introduction

Following the same line of the previous two appendices, we present the Stata features to conduct a complete data analysis, starting from the data loading until the exporting of the results. Also in this case we assume the data are stored in the file *example. {txt, csv, xls, dta, sav}*. The file extension denotes the format of the data file, as follows:

txt tab delimited text file;
csv comma separated values file;
xls spreadsheet (Microsoft Excel) file;
dta Stata file;
sav Spss data file.

The variables stored in the file will be denoted as follows:

y dependent variable;
$x1$**,** $x2$**,** $x3$**,** $x4$ set of explanatory numerical variables;
$z1, z2$ set of explanatory categorical variables.

Throughout the appendix, Stata commands are shown in bold font and comments using regular font. The commands shown are fully explained in the Stata official documentation (Stata 2011a; Stata 2011c). Among the different books dealing with Stata, the books by Acock (2012), Hamilton (2012), and Scott Long (2008) offer a complete description of the use of the software for carrying out a statistical analysis.

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

C.1 Loading data

The most common formats for text data are tab delimited files and comma separated values files. The typical extension of the former files is '.txt' while for the latter the '.csv' extension is generally used. We assume the data file is in the current working directory.

```
// to obtain the current working directory
pwd

// to set the working directory by specifying the
// complete path
// in the example the working directory is set to the
// bookQR folder contained in the folder Documents
cd "c:\Documents\bookQR"
```

C.1.1 Text data

The *infile* command allows to import an ASCII file where the columns can be delimited by different characters.

```
// the infile command reads into memory an ASCII file
// in which the values are separated by one or more
// whitespace characters or by commas
// the str# statement precedes the string variable z1
// and z2 the number # denotes the string length
infile y x1 x2 x3 x4 str30 z1 str30 z2 using
       example.txt, clear
// to automatically change all variables to their most
// memory-efficient storage type
compress
describe

// the option automatic allows to manage non-numeric
// variables so to use strings as labels for the
// levels of the categorical variables
infile y x1 x2 x3 x4 z1 z2 using example.txt, automatic
```

An alternative command is the *insheet* command, that allows to specify the column delimiter as well as to manage the name of variables if present on the first row of the file to import.

```
// the insheet command allows to specify the column
// delimiter
// the tab option is used for tab delimited files
insheet y x1 x2 x3 x4 z using example.txt, automatic tab

// the comma option is used for comma separated files
insheet y x1 x2 x3 x4 z using example.txt, automatic
        comma

// the option names is used to denote that the
// variables' names are on the first row
// the variables' list can be omitted
insheet using example.txt, automatic tab names
```

C.1.2 Spreadsheet data

While in the past the official Stata documentation tooltip for reading Excel files consisted of first saving them as comma separated values (.csv) files and then using the *insheet* command, starting from *Stata 12* the *import excel* command allows Excel (.xls and .xlsx) files to be read directly.

```
// the import excel command allows to directly read
// spreadsheets to use
// to read the first sheet of the
// workbook whereas there are not column names on
// the first row
import excel y x1 x2 x3 x4 z example.xls, clear

// the firstrow option allows to indicate that the
// first row contains the variable names
import excel example.xls, firstrow clear

// the sheet option indicates the name of the worksheet
// hosting the data to import
import excel using hospital.xls, sheet(Sheet1) clear

// it is possible to specify a range of cells to import
// using a spreadsheet-like syntax exploiting the
// cellrange option
import excel example.xls, firstrow cellrange(A1:F101)
clear
```

C.1.3 Files from other statistical packages

The *Stat/Transfer* and *DBMS/Copy* softwares offer specialized conversion tools to convert files in different formats to Stata. It is also possible to import files in different formats exploiting an ODBC driver, whereas a driver for the required software is correctly installed and configured.

```
// to load a data stored in a stata format file
use example.dta, clear

// to load a Spss data file
// it does not require any other software installed
// on the PC
usespss example.sav, clear
// convert the Spss format to a stata native format
usespss example.sav, saving(example.dta)

// to import a SAS Xport format file
fdause example.xpt, clear
// to import a SAS data file
// it requires SAS to be installed on the PC
usesas example.ssd, clear
```

C.2 Exploring data

This section shows some of the main commands for exploring data through graphical representation and summary tables. Although in the text we refer to the commands with respect to the dependent variable y, they can be useful for any numerical variable. The choice to refer to the dependent variable is due to the typical approach in a regression analysis, where the study of the y characteristics can suggest the best model to use.

C.2.1 Graphical tools

The graphical commands shown in this section are detailed in the Stata Graphics documentation (Stata 2011b). An additional useful reading for the different representations available in Stata is the book written by Mitchell (2012), which offers a visual tour of the different Stata graphical tools.

C.2.1.1 Histogram of the dependent variable

```
// to draw histogram of the y variable using frequencies
// on the vertical axis
```

```
histogram y, frequency

// the width option allows to specify the width of
// the bins
histogram y, width(5) frequency

// to draw histogram of the y variable using fractions
// of the data on the vertical axis
histogram y, fraction

// to draw histogram of the y variable using fractions
// of the data on the vertical axis and adding a
// normal curve based on the sample mean and sample
// standard deviation
histogram y, norm fraction
```

C.2.1.2 Conditional histograms of the dependent variable

```
// to draw separate histograms of the y variable for
// each value of the z1 variable, showing the fraction
// of the data on the vertical axis
histogram y, by(z1) frequency

// to draw separate histograms of the y variable for
// each value of the z1 variable along with a total
// histogram for the whole sample
histogram y, by(z1, total) frequency

// to draw separate histograms of the y variable for
// each value of the cartesian product between
// the variables z1 and z2
egen group = group(z1 z2), label
histogram y, by(group) frequency
```

C.2.1.3 Boxplot of the dependent variable

```
// boxplot of the y variable
graph box y

// boxplot of a set of variables
graph box y x1 x2 x3 x4
```

C.2.1.4 Conditional boxplots of the dependent variable

```
// boxplot of the y variable for each values of the
// z1 variable
graph box y, over(z1)

// to draw separate boxplots of the y variable for
// each value of the cartesian product between the z1
// and z2 variables
egen group = group(z1 z2), label
graph box y, by(group) frequency
```

C.2.1.5 Quantile plot of the dependent variable

```
// normal probability plot: quantiles of y vs
// corresponding quantiles of a normal distribution
qnorm y
```

C.2.1.6 Conditional quantile plots of the dependent variable

```
// normal probability plot: quantiles of y vs
// corresponding quantiles of a normal distribution for
// each value of z1 variable
qnorm y, by(z1)
```

C.2.1.7 Scatter plot

```
// to display a scatterplot of x1 vs y
twoway (scatter y x1)
```

C.2.1.8 Conditional scatter plots

```
// to display a scatterplot of x1 vs y
// using colours for the different levels
// of a categorical variable (z1)
// we suppose that z1 has the following
// three levels: "lev1", "lev2", "lev3"
twoway (scatter y x if z1=="lev1",
                mcolor(red) msymbol(circle_hollow))
       (scatter y x if z1=="lev2",
          mcolor(midgreen) msymbol (triangle_hollow))
       (scatter y x if z1=="lev3",
                mcolor(blue) msymbol(plus))
```

```
// using different panels for the levels of a single
// categorical variable
twoway (scatter y x1), by(z1)

// using different panels for the levels of the
// cartesian product between the variables z1 and z2
egen group = group(z1 z2), label
twoway (scatter y x1), by(group)
```

C.2.2 Summary statistics

In this section the commands for obtaining the main summary statistics are reported. In particular, we present univariate summary statistics and synthesis tables for one and two variables.

C.2.2.1 Summary statistics of the dependent variable

```
// summary statistics (means, standard deviations,
// minimum and maximum values, and number of
// observations) for the listed variable
summarize y
// detailed summary statistics, including percentiles,
// median, mean, standard deviation, variance, skewness
// and kurtosis
summarize y, details

// detailed summary statistics for a set of variables
summarize y x1 x2, details

// specified summary statistics for a single variable
tabstat y, stats(mean sd skewness kurtosis)

// location indexes for a single variable
tabstat y, stats(min p5 p25 p50 p75 p95 max)
```

C.2.2.2 Conditional summary statistics of the dependent variable

```
// summary statistics (means, standard deviations,
// minimum and maximum values, and number of observation)
```

```
// for the listed variable within categories of the
// z1 variable
summarize y, by(z1)

// detailed summary statistics, including percentiles,
// median, mean, standard deviation, variance, skewness
// and kurtosis within categories of the z1 variable
summarize y, details by(z1)

// specified summary statistics for a single variable
// within categories of the z1 variable
tabstat y, stats(mean sd skewness kurtosis) by(z1)

// location indexes for a single variable
// within categories of the z1 variable
tabstat y, stats(min p5 p25 p50 p75 p95 max) by(z1)

// location indexes for a single variable
// within categories of the cartesian product
// between the variables z1 and z2
egen group = group(z1 z2), label
tabstat y, stats(mean sd skewness kurtosis) by(group)
```

C.2.2.3 Contingency tables

```
// frequency table for a categorical variable
tabulate z1

// contingency table for two categorical variables
tabulate z1 z2
```

C.3 Modeling data

C.3.1 Ordinary least squares regression analysis

C.3.1.1 Parameter estimates

```
// OLS estimation
regress y x1

// OLS fitted values
```

```
regress y x1
predict yfit

// OLS residuals
regress y x1
predict res, residuals

// regression introducing a categorical variable
// the option gen of the tabulate command generates
// a dummy variable for each level of the variable
tabulate z1, gen(dumZ)
// assuming that the z1 variable has three levels
describe dumZ1-dumZ3
// regression using the generated dummy variables
regress y dumZ1 dumZ2 dumZ3

// multiple regression model
regress y x1 x2 x3 x4
```

C.3.1.2 Main graphical representations

```
// to draw a residual versus fitted plot, using the
// most recent regression
rvfplot
// residual versus fitted plot, adding an horizontal
// line at y=0
rvfplot, yline(0)

// residuals versus the values of the predictor x1
rvpplot x1

// to draw the simple regression line
graph twoway lfit y x ||  scatter y x

// scatterplot of the residuals
twoway (scatter res x)

// component-plus-residual-plot useful for screening
// nonlinearities
cprplot x1

// augmented component-plus-residual-plot useful for
// screening nonlinearities
```

```
acprplot x1

// added-variable plot useful for detecting
// influence points
avplot x1

// combines in one image all the added-variable plots
// from the most recent regress command
avplots

// leverage-versus-squared-residual plot
lvr2plot
```

C.3.1.3 Confidence intervals and hypothesis tests

```
// confidence intervals and hypothesis tests are
// built in results
// however the vector of estimated coefficients and
// the variance covariance matrix can be saved and
// printed by adding the following instructions
// immediately after the regression instruction
reg y x1 x2
matrix coefname=e(b)
matrix list coefname
matrix varname=e(V)
matrix list varname

// there are additional elements that can be saved and
// printed, like the explained sum of squares and the
// residuals sum of squares
scalar modelname=e(mss)
scalar list modelname
scalar residualname=e(rss)
scalar list residualname
```

C.3.2 Quantile regression analysis

C.3.2.1 Parameter estimates

```
// QR estimation: inside parenthesis is the
// selected quantile (in this case the only median)
qreg y x, q(.5)

// variance covariance matrix
```

```
matrix var=e(V)
matrix list var

// QR fitted values
predict qfit

// QR residuals
predict qres

// QR estimation for more quantiles
qreg y x, q(.10 .25 .5 .75 .9)
```

C.3.2.2 Main graphical representations

In order to obtain a graphical representation of the QR coefficients, it is possible to exploit the *GRQREG* module (Azevedo 2011).

```
// to install the grqreg module
ssc install grqreg
// after the installation, the grqreg command allows
// to plot the QR coefficients
// it works after the commands: qreg, bsqreg, sqreg
// it has the option to graph the confidence interval,
// the OLS coefficient and the OLS confidence interval
// on the same graph

// QR estimation
qreg y x
// QR coefficient plot for the slope
// by default the graph for all the estimated
// coefficients except the intercept are produced
grqreg
// QR coefficient plot for the intercept
grqreg, cons

// to set the minimum and maximum values, and the
// steps for the quantiles
// minimum (qmin) default = .05
// maximum (qmax) default = .95
// increment (qstep) default = .05
gqreg y x, qmin = .01 qmax=.99 qstep=.01
// to draw the QR confidence intervals
gqreg, ci level=0.05

// to draw the OLS line, the OLS confidence intervals
// along with their QR counterpart
gqreg, ols olsci ci level=0.05
```

In the case of a multiple QR, the *gqreg* command allows to specify the coefficients to plot through the *varlist* option. It supports also the comparison of two coefficients exploiting the *compare* option.

```
// QR multiple regression
qreg y x1 x2 x3 x4

// graph for all the estimated coefficients
// except the intercept
grqreg

// graph for all the estimated coefficients
// along with the intercept
grqreg, cons

// graph for the specified coefficients
// along with the intercept
grqreg x1 x4, cons

// the compare option graphs two QR coefficients
// it supports two variables at the time: the first
// variable is plotted on the y-axis and the second
// variable on the x-axis
grqreg  x1 x2, compare
```

```
// scatter plot of the residuals
twoway (scatter qres x)

// scatter plot of the data together with OLS and QR
// estimated regressions
// yfit are the OLS fitted values and qfit are the
// QR fitted values
// these values are computed by the OLS and the QR
// regressions
twoway (scatter y x) (line yfit x) (line qfit x)
```

C.3.2.3 Bootstrap estimates

```
// to compute 100 replicates of 25-th quantile
// regression coefficients, and save the estimates to
// the file outBootstrap
bootstrap, qreg y x1, q(.25) _b, reps(100)
sa(outBootstrap)
```

C.3.2.4 Confidence intervals and hypothesis tests

```
// confidence intervals and hypothesis tests are
// built in results
// however the vector of estimated coefficients and the
// variance covariance matrix can be saved and printed
// by adding the following instructions immediately
// after the regression instruction
// these values are computed by the OLS and the QR
// regressions
qreg y x z1, q(.9)
matrix coefnameqr=e(b)
matrix list coefnameqr
matrix varnameqr=e(V)
matrix list varnameqr

// there are additional elements that can be saved and
// printed, like the height of the error density at
// the selected quantile and the estimated objective
// function
// the error density at the specified quantile is
// relevant to compute sparsity and scale
scalar densityname=e(f_r)
scalar list densityname
scalar objectivename=e(sum_adev)
scalar list objectivename

// to test more than one exclusion restriction
// unconstrained median regression
qreg y x1 x2 x3 x4, q(.5)

// to verify the null H0:b2=b3=0
test x2 x3
// to verify the null H0:b2=b3=b4=0
test x2 x3 x4

// test for QR heteroskedasticity
// for five conditional quantiles
sqreg y x1, q(.10 .25 .5 .75 .9)

// test for the equality of two slopes: 10-th conditional
// quantile vs 90-th conditional quantile
test [q10]x1 = [q90]x1
// it is possible to test the equality of two slopes
// also exploiting the lincom command
```

```
lincom [q90]x1-[q10]x1

// the test of equality between two slopes can also be
// directly obtained using the iqreg command
iqreg y x1, q(.10, .90)

// use the option accum of the command test
// to test the equality of more than two slopes
// together
sqreg y x1, q(.10 .25 .5 .75 .9)
test [q25]x1 = [q10]x1
test [q5]x1 = [q25]x1, accum
test [q75]x1 = [q5]x1, accum
test [q9]x1 = [q75]x1, accum
```

C.4 Exporting figures and tables

C.4.1 Exporting figures

```
// to save a graph in stata graphic format
graph save graph1.gph

// to export a graph in pdf format
graph export graph1.pdf, as(pdf)
//NOTE: main available format
//  ps: Adobe PostScript file
// eps: Encapsuled PostScript
// wmf: Windows Metafile
// emf: Windows Enhanced Metafile
// png: PNG bitmap file
// tif: TIFF file
```

C.4.2 Exporting tables

```
// to compute the 10-th conditional quantile
qreg y x z1, q(.1)

// to save in tablename coefficients and standard errors
outreg using tablename, se ct(.10) replace
//se=standard error, ct=column name in the table
outreg using tablename, se ct(OLS) append
```

```
// NOTE 1:
// the output file is tab-delimited and has the
// extension .out
// open the file in Word, highlight the rows of
// results, click on Insert, Table, Convert
// text to table.

// NOTE 2:
// the outreg2 command is an extension of the outreg
// command in order to enhance its capabilities with
// multiple models and to provide more format options
```

Among the other available commands for exporting tables, the following are worth mentioning:

xmlsave allows to export data to Microsoft Excel exploiting an XML format;

tabout produces publication–quality cross tabulations, allowing to export the table both in tab-delimited and in HTML format;

xml_tab converts stored estimates in Excel's XML format;

logout captures the results that are printed to the Results Window (log) and writes them to Excel, Word, or other formats.

References

Acock AC 2012 *A Gentle Introduction to Stata*. Stata Press.

Azevedo JP 2011 *GRQREG: Stata Module to Graph the Coefficients of a Quantile Regression*. http://econpapers.repec.org/software/bocbocode/s437001.htm (accessed April 2013).

Hamilton LW 2012 *Statistics with Stata: Version 12*. Duxbury Press.

Mitchell N 2012 *A Visual Guide to Stata Graphics*. Stata Press.

Scott Long J 2008 *The Workflow of Data Analysis with Stata*. Stata Press.

Stata 2011a *Stata 12. Base Reference Manual (4 volumes)*. Stata Press.

Stata 2011b *Stata 12. Graphics Reference Manual*. Stata Press.

Stata 2011c *Stata 12. User's Guide*. Stata Press.

Index

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

WILEY SERIES IN PROBABILITY AND STATISTICS
ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Second Edition*
ALSTON, MENGERSEN and PETTITT (editors) · Case Studies in Bayesian Statistical Modelling and Analysis
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

BARTHOLOMEW, KNOTT, and MOUSTAKI · Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*

BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference, *Second Edition*

BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications

BATES and WATTS · Nonlinear Regression Analysis and Its Applications

BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL · Statistics of Extremes: Theory and Applications

BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression

† BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*

BERNARDO and SMITH · Bayesian Theory

BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*

BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications

BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys

BILLINGSLEY · Convergence of Probability Measures, *Second Edition*

BILLINGSLEY · Probability and Measure, *Anniversary Edition*

BIRKES and DODGE · Alternative Methods of Regression

BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example

BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics

BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance

BLISCHKE and MURTHY · Reliability: Modeling, Prediction, and Optimization

BLOOMFIELD · Fourier Analysis of Time Series: An Introduction,

Second Edition

BOLLEN · Structural Equations with Latent Variables

BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective

BOROVKOV · Ergodicity and Stability of Stochastic Processes

BOSQ and BLANKE · Inference and Prediction in Large Dimensions

BOULEAU · Numerical Methods for Stochastic Processes

* BOX · Bayesian Inference in Statistical Analysis

BOX · Improving Almost Anything, *Revised Edition*

* BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement

BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses,

Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Editon*

BOX, JENKINS, and REINSEL · Time Series Analysis: Forcasting and Control, *Fourth Edition*

BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*

* BROWN and HOLLANDER · Statistics: A Biomedical Introduction

CAIROLI and DALANG · Sequential Stochastic Optimization

CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science

CHAN · Time Series: Applications to Finance with R and

S-Plus®, *Second Edition*

CHARALAMBIDES · Combinatorial Methods in Discrete Distributions

CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*

CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression

CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*

CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences

CHILES and DELFINER · Geostatistics: Modeling Spatial Uncertainty,

Second Edition

CHIU, STOYAN, KENDALL and MECKE · Stochastic Geometry and its Applications, *Third Edition*

CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*

CLARKE · Linear Models: The Theory and Application of Analysis of Variance

CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*

* COCHRAN and COX · Experimental Designs, *Second Edition*

COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences

CONGDON · Applied Bayesian Modelling

CONGDON · Bayesian Models for Categorical Data

CONGDON · Bayesian Statistical Modelling, *Second Edition*

CONOVER · Practical Nonparametric Statistics, *Third Edition*

COOK · Regression Graphics

COOK and WEISBERG · An Introduction to Regression Graphics

COOK and WEISBERG · Applied Regression Including Computing and Graphics

CORNELL · A Primer on Experiments with Mixtures

CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*

COX · A Handbook of Introductory Statistical Methods

CRESSIE · Statistics for Spatial Data, *Revised Edition*

CRESSIE and WIKLE · Statistics for Spatio-Temporal Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis

DAGPUNAR · Simulation and Monte Carlo: With Applications in Finance and MCMC

DANIEL · Applications of Statistics to Industrial Experimentation

DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*

* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*

DASU and JOHNSON · Exploratory Data Mining and Data Cleaning

DAVID and NAGARAJA · Order Statistics, *Third Edition*

DAVINO, FURNO and VISTOCCO · Quantile Regression: Theory and Applications

* DEGROOT, FIENBERG, and KADANE · Statistics and the Law

DEL CASTILLO · Statistical Process Adjustment for Quality Control

DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables

DEMIDENKO · Mixed Models: Theory and Applications

DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression

DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis

DEY and MUKERJEE · Fractional Factorial Plans

DE ROCQUIGNY · Modelling Under Risk and Uncertainty: An Introduction to Statistical, Phenomenological and Computational Models

DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications

* DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*

* DOOB · Stochastic Processes

DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*

DRAPER and SMITH · Applied Regression Analysis, *Third Edition*

DRYDEN and MARDIA · Statistical Shape Analysis

DUDEWICZ and MISHRA · Modern Mathematical Statistics

DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*

DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations

EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment

* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis

ENDERS · Applied Econometric Time Series, *Third Edition*

† ETHIER and KURTZ · Markov Processes: Characterization and Convergence

EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*

EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, *Fifth Edition*

FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs

FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*

FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

* FLEISS · The Design and Analysis of Clinical Experiments

FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*

† FLEMING and HARRINGTON · Counting Processes and Survival Analysis

FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations

FULLER · Introduction to Statistical Time Series, *Second Edition*

† FULLER · Measurement Error Models

GALLANT · Nonlinear Statistical Models

GEISSER · Modes of Parametric Statistical Inference

GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives

GEWEKE · Contemporary Bayesian Econometrics and Statistics

GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation

GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments

GIFI · Nonlinear Multivariate Analysis

GIVENS and HOETING · Computational Statistics

GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems

GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*

GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*

GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues

GOLDSTEIN and WOOFF · Bayes Linear Statistics

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*

GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*

* HAHN and SHAPIRO · Statistical Models in Engineering

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

HALD · A History of Probability and Statistics and their Applications Before 1750

† HAMPEL · Robust Statistics: The Approach Based on Influence Functions

HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications

HEIBERGER · Computation for the Analysis of Designed Experiments

HEDAYAT and SINHA · Design and Inference in Finite Population Sampling

HEDEKER and GIBBONS · Longitudinal Data Analysis

HELLER · MACSYMA for Statisticians

HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special Designs and Applications

* HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance

* HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes

* HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis

HOCHBERG and TAMHANE · Multiple Comparison Procedures

HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*

HOEL · Introduction to Mathematical Statistics, *Fifth Edition*

HOGG and KLUGMAN · Loss Distributions

HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*

HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*

HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*

HUBER · Data Analysis: What Can Be Learned From the Past 50 Years

HUBER · Robust Statistics

† HUBER and RONCHETTI · Robust Statistics, *Second Edition*

HUBERTY · Applied Discriminant Analysis, *Second Edition*

HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*

HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*

HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*

HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice

HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary

HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

INSUA, RUGGERI and WIPER · Bayesian Analysis of Stochastic Process Models

JACKMAN · Bayesian Analysis for the Social Sciences

† JACKSON · A User's Guide to Principle Components

JOHN · Statistical Methods in Engineering and Quality Assurance

JOHNSON · Multivariate Statistical Simulation

JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz

JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*

JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*

JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions

JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*

JUREK and MASON · Operator-Limit Distributions in Probability Theory

KADANE · Bayesian Methods and Ethics in a Clinical Trial Design

KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence

KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*

KARIYA and KURATA · Generalized Least Squares

KASS and VOS · Geometrical Foundations of Asymptotic Inference

† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis

KEDEM and FOKIANOS · Regression Models for Time Series Analysis

KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory

KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*

KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models

* KISH · Statistical Design for Research

KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences

KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization

KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*

KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*

KOSKI and NOBLE · Bayesian Networks: An Introduction

KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume

KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1

KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2

KOWALSKI and TU · Modern Applied U-Statistics

KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation

KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods

KROONENBERG · Applied Multiway Data Analysis

KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory

KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling

KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering

LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*

LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction

LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*

LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*

LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*

LE · Applied Categorical Data Analysis, *Second Edition*

LE · Applied Survival Analysis

LEE · Structural Equation Modeling: A Bayesian Approach

LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*

LEPAGE and BILLARD · Exploring the Limits of Bootstrap

LESSLER and KALSBEEK · Nonsampling Errors in Surveys

LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics

LIAO · Statistical Group Comparison

LIN · Introductory Stochastic Analysis for Finance and Insurance

LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*

LLOYD · The Statistical Analysis of Categorical Data

LOWEN and TEICH · Fractal-Based Point Processes

MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*

MALLER and ZHOU · Survival Analysis with Long Term Survivors

MARCHETTE · Random Graphs for Statistical Pattern Recognition

MARDIA and JUPP · Directional Statistics

MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice

MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods

MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*

McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*

McFADDEN · Management of Data in Clinical Trials, *Second Edition*

* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition

McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data

McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*

McLACHLAN and PEEL · Finite Mixture Models

McNEIL · Epidemiological Research Methods

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

MEEKER and ESCOBAR · Statistical Methods for Reliability Data

MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice

MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and Applications

MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*

* MILLER · Survival Analysis, *Second Edition*

MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting

MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fifth Edition*

MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness

MUIRHEAD · Aspects of Multivariate Statistical Theory

MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks

MURTHY, XIE, and JIANG · Weibull Models

MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*

MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*

NATVIG · Multistate Systems Reliability Theory With Applications

† NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses

† NELSON · Applied Life Data Analysis

NEWMAN · Biostatistical Methods in Epidemiology

NG, TAIN, and TANG · Dirichlet Theory: Theory, Methods and Applications

OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, *Second Edition*

OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis

PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions

PANJER · Operational Risk: Modeling and Analytics

PANKRATZ · Forecasting with Dynamic Regression Models

PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases

PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance

PARMIGIANI and INOUE · Decision Theory: Principles and Approaches

* PARZEN · Modern Probability Theory and Its Applications

PEÑA, TIAO, and TSAY · A Course in Time Series Analysis

PESARIN and SALMASO · Permutation Tests for Complex Data: Applications and Software

PIANTADOSI · Clinical Trials: A Methodologic Perspective, *Second Edition*

POURAHMADI · Foundations of Time Series Analysis and Prediction Theory

POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

POWELL and RYZHOV · Optimal Learning

PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*

PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach

PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics

† PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming

QIU · Image Processing and Jump Regression Analysis

* RAO · Linear Statistical Inference and Its Applications, *Second Edition*

RAO · Statistical Inference for Fractional Diffusion Processes

RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*

RAYNER, THAS, and BEST · Smooth Tests of Goodnes of Fit: Using R, *Second Edition*

RENCHER and SCHAALJE · Linear Models in Statistics, *Second Edition*

RENCHER and CHRISTENSEN · Methods of Multivariate Analysis, *Third Edition*

RENCHER · Multivariate Statistical Inference with Applications

RIGDON and BASU · Statistical Methods for the Reliability of Repairable Systems

* RIPLEY · Spatial Statistics

* RIPLEY · Stochastic Simulation

ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*

ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance

ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice

ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing

† ROUSSEEUW and LEROY · Robust Regression and Outlier Detection

ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables

* RUBIN · Multiple Imputation for Nonresponse in Surveys

RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*

RUBINSTEIN and MELAMED · Modern Simulation and Modeling

RYAN · Modern Engineering Statistics

RYAN · Modern Experimental Design

RYAN · Modern Regression Methods, *Second Edition*

RYAN · Sample Size Determination and Power

RYAN · Statistical Methods for Quality Improvement, *Third Edition*

SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications

SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis

SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions

* SCHEFFE · The Analysis of Variance

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application

SCHOTT · Matrix Analysis for Statistics, *Second Edition*

SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives

SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization

* SEARLE · Linear Models

† SEARLE · Linear Models for Unbalanced Data

† SEARLE · Matrix Algebra Useful for Statistics

† SEARLE, CASELLA, and McCULLOCH · Variance Components

SEARLE and WILLETT · Matrix Algebra for Applied Economics

SEBER · A Matrix Handbook For Statisticians

† SEBER · Multivariate Observations

SEBER and LEE · Linear Regression Analysis, *Second Edition*

† SEBER and WILD · Nonlinear Regression

SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems

* SERFLING · Approximation Theorems of Mathematical Statistics

SHAFER and VOVK · Probability and Finance: It's Only a Game!

SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties

SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions

SINGPURWALLA · Reliability and Risk: A Bayesian Perspective

SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference

SRIVASTAVA · Methods of Multivariate Statistics

STAPLETON · Linear Statistical Models, *Second Edition*

STAPLETON · Models for Probability and Statistical Inference: Theory and Applications

STAUDTE and SHEATHER · Robust Estimation and Testing

STOYAN · Counterexamples in Probability, *Second Edition*

STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*

STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics

STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods

STYAN · The Collected Papers of T. W. Anderson: 1943–1985

SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research

TAKEZAWA · Introduction to Nonparametric Regression

TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications

TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory

THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*

THOMPSON · Sampling, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

THOMPSON · Simulation: A Modeler's Approach

THOMPSON and SEBER · Adaptive Sampling

THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets

TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics

TSAY · Analysis of Financial Time Series, *Third Edition*

UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data

† VAN BELLE · Statistical Rules of Thumb, *Second Edition*

VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*

VESTRUP · The Theory of Measures and Integration

VIDAKOVIC · Statistical Modeling by Wavelets

VIERTL · Statistical Methods for Fuzzy Data

VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments

WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data

WANG and WANG · Structural Equation Modeling: Applications Using M*plus*

WEISBERG · Applied Linear Regression, *Third Edition*

WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons

WELSH · Aspects of Statistical Inference

WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment

* WHITTAKER · Graphical Models in Applied Multivariate Statistics

WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting

WOODWORTH · Biostatistics: A Bayesian Introduction

WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*

WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*

WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis

YIN · Clinical Trial Design: Bayesian and Frequentist Adaptive Methods

YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics

ZACKS · Stage-Wise Adaptive Designs

* ZELLNER · An Introduction to Bayesian Inference in Econometrics

ZELTERMAN · Discrete Distributions—Applications in the Health Sciences

ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.