Handling heterogeneity in Quantile Regression

Cristina Davino

Department of Economics and Statistics
University of Naples Federico II
Italy
`cristina.davino@unina.it`

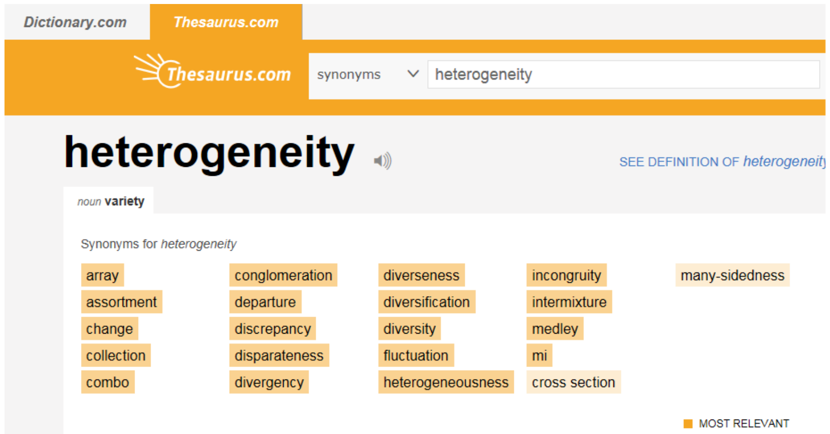
Cnam - Paris
April 5, 2019

Outline

all computations and graphics were done in the R language using the packages *quantreg* and *plspm*

Heterogeneity

Synonyms for heterogeneity



Heterogeneity

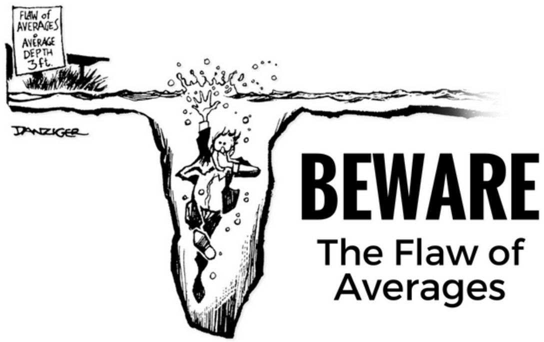
High heterogeneity is often more realistic for modeling the messy real world and may give better results or identify subpopulations

Part 1: Quantile Regression

Motivation

(Koenker R W and Basset G, Regression Quantiles. *Econometrica* **46**(1), 1978)

The flaw of Averages: a rationale for quantile regression



Basic motivation

Mosteller and Tukey (1977)

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of **X**'s.
We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set.
Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

Quantile regression

- QR has become a popular alternative to least squares regression for modeling heterogeneous data
- QR gained popularity in applied economics by the end of the 90's, when people realize the importance of heterogeneity
- Fields of application:
  - astrophysics
  - chemistry
  - ecology
  - economics
  - finance
  - genomics
  - medicine
  - meteorology
  - sociology
  - marketing
  - food science

Classical linear regression

Classical linear regression (conditional expected value)

estimation of the conditional mean of a response variable (y) distribution as a function of a set X of predictor variables

Pros

- gives a parsimonious description of the dependent relationship
- estimators with several properties
- ...

Cons:

- Heteroscedastic relationships
- Presence of outliers
- Skewed dependent variable

Classical vs quantile linear regression

Classical linear regression (conditional expected value)

estimation of the conditional mean of a response variable (y) distribution as a function of a set X of predictor variables

$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta$

Quantile regression (conditional quantiles)

estimation of the conditional quantiles of a response variable (y) distribution as a function of a set X of predictor variables

$Q_\theta(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta(\theta)$

where: $(0 < \theta < 1)$

(Koenker R., Basset G. 1978) (Koenker R. 2005)
(Koenker R. `quantreg R package` 2018)
(Davino C., Furno M., Vistocco D. 2013) (Furno M., Vistocco D. 2018)

Quantile Regression model

QR model for a given conditional quantile θ (linear regression):

$$Q_\theta(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta(\theta)$$

where

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

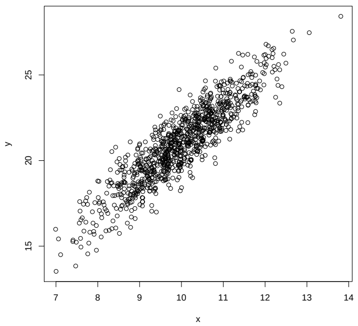
- Classical regression focuses on $E(\mathbf{y}|\mathbf{X})$
- QR extends this approach to study the conditional distribution of a response variable
- θ regression lines are estimated
- The estimation of coefficients for each quantile regression is based on the whole sample, not just the portion of the sample at that quantile

Two examples with simulated data

homogeneous model

$\mathbf{y}_1 = 1 + 2\mathbf{x} + \mathbf{e}$

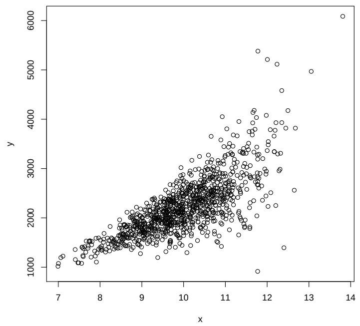
$\mathbf{x} \sim N(10;1) \quad \mathbf{e} \sim N(0;1)$



heterogeneous model

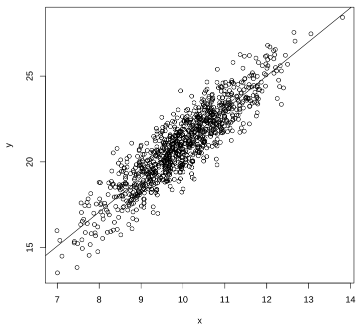
$\mathbf{y}_2 = 1 + 2\mathbf{x} + (1+\mathbf{x})\mathbf{e}$

$\mathbf{x} \sim N(10;1) \quad \mathbf{e} \sim N\left((-1+20x); e^{x/3}\right)$

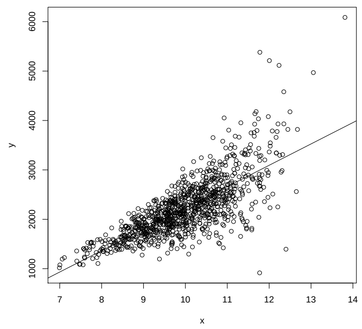


OLS results

homogeneous model

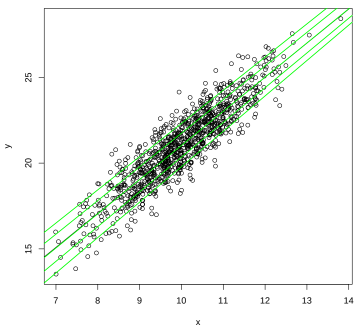


heterogeneous model

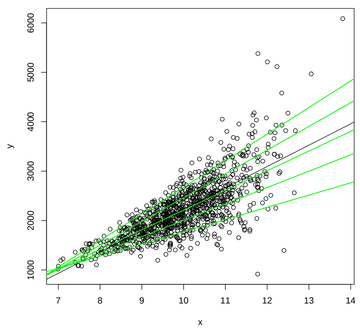


OLS and QR results

homogeneous model



heterogeneous model



OLS and QR results

homogeneous model

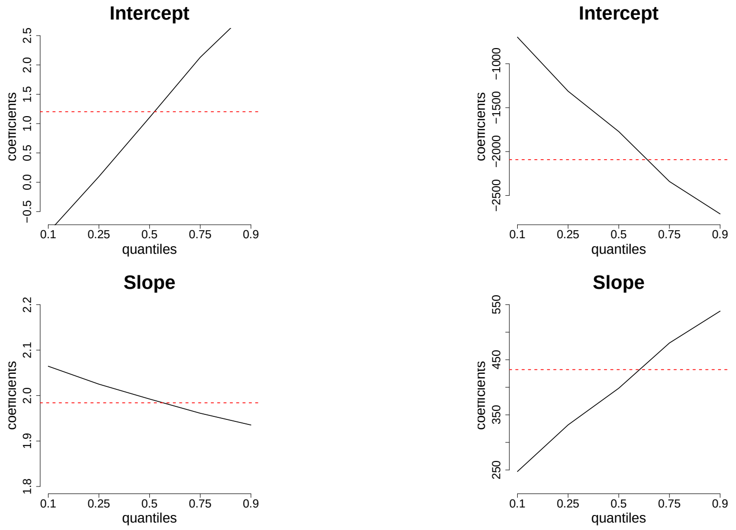
	OLS	θ= 0.1	θ= 0.25	θ= 0.5	θ= 0.75	θ= 0.9
intercept	0.5	-0.5	-0.7	0.4	1.6	1.2
x	2.0	2.0	2.1	2.1	2.0	2.1

heterogeneous model

	OLS	θ= 0.1	θ= 0.25	θ= 0.5	θ= 0.75	θ= 0.9
intercept	-2092.0	-697.2	-1312.7	-1772.2	-2340.6	-2709.7
x	432.1	247.1	331.8	398.3	480.4	538.3

OLS and QR results

homogeneous model *heterogeneous model*



Quantile Regression model

Interpretation

$$\hat{\beta}_i(\theta) = \frac{\partial Q_\theta(\mathbf{y}|\mathbf{X})}{\partial \mathbf{x}_i}$$

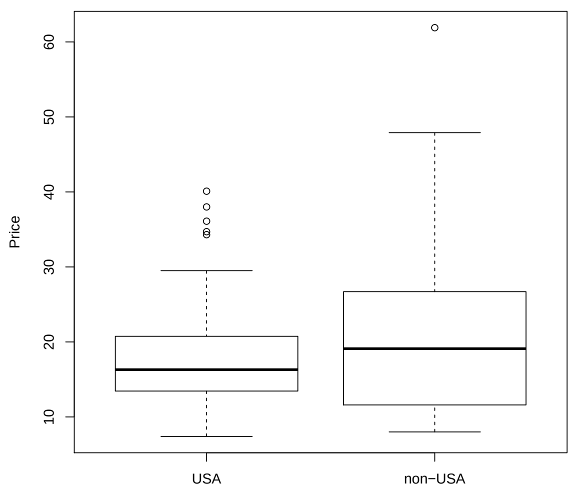
- Rate of change of the θ^{th} quantile of the dependent variable per unit change in the value of the i^{th} quantile
- Fitted values reconstruct the conditional quantiles
- QR generalizes univariates quantiles for conditional distributions

QR pros:

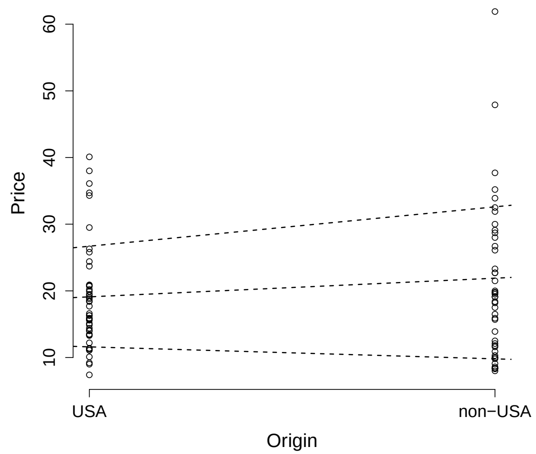
- Regressor effects on the whole dependent variable distribution
- Heteroscedastic relationships
- Presence of outliers
- Skewed dependent variable

A simple example: the ‘93cars’ dataset

- 93 new cars for the 1993 model year
- selected measures: Price, Origin (USA , non-USA), Horsepower



A simple example: the ‘93cars’ dataset



- Slopes: rate of change of the y θ^{th} conditional quantile per unit change of the regressor
- Fitted values reconstruct the conditional quantiles
- QR generalizes univariates quantiles for conditional distributions

	USA	non-USA	uncond.	intercept	slope
Mean	18.6	20.5	19.5	18.6	1.9
θ=0.25	13.5	11.6	12.2	13.4	-1.8
θ=0.5	16.3	19.1	17.7	16.3	2.8
θ=0.75	20.7	26.7	23.3	20.8	5.9

Source

The quantile process and the selection of the quantiles

- QR solutions are typically computed for a selected number of quantiles
- It is possible to obtain estimates across the entire interval of conditional quantiles
- A dense grid of equally spaced quantiles provides a fairly accurate approximation of the whole quantile regression pattern
- The number of distinct quantiles is related to: the number of units and the number of variables

Part 1: Quantile Regression

Estimation

Unconditional mean and quantiles

QR is to classical regression what quantiles are to mean in terms of describing locations of a distribution

Let Y be a generic random variable:

- Mean (and its objective function): $\mu = \arg\min_c E(Y - c)^2$
- Median (and its objective function): $Me = \arg\min_c E|Y - c|$
- Generic quantile θ (and its objective function):

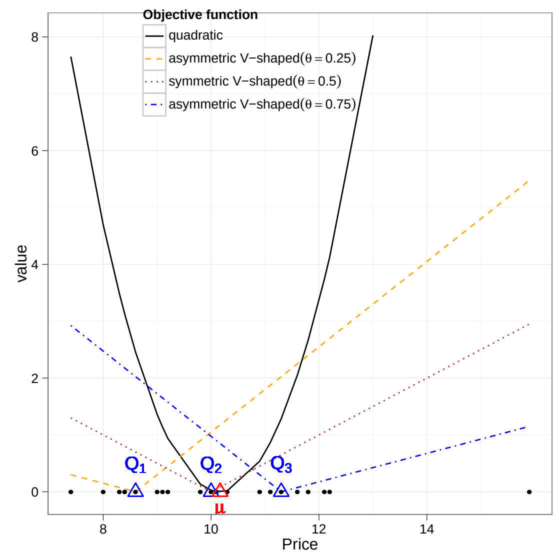
$$q_\theta = \arg\min_c E[\rho_\theta(Y - c)]$$

- $\hat{\mu}$ e $\hat{Me}$ denotes the sample estimators for such centers
- $\rho_\theta(.)$ denotes the following location functions:

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

- $\rho_\theta(.)$ is 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

On optimal criteria



Conditional mean and conditional quantiles estimation

Least squares linear regression estimator

$$\hat{\beta} = \arg\min_{\beta} E\,[\mathbf{y} - \mathbf{X}\beta)]^2$$

Conditional quantile linear regression estimator

$$\hat{\beta}(\theta) = \arg\min_{\beta} E\,[\rho_\theta(\mathbf{y} - \mathbf{X}\beta)]$$

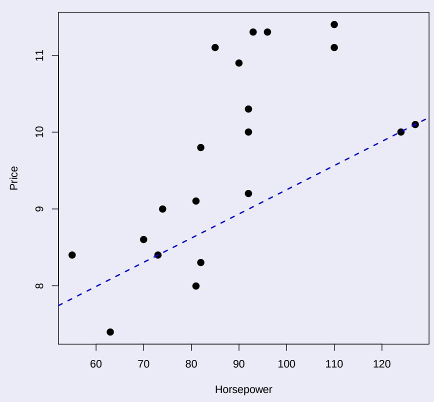
Note: the (θ)–notation denotes that the parameters and the corresponding estimators are for a specific quantile θ

$\rho_\theta(.)$ is 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.

$$\rho_\theta = \begin{cases} \theta\,(u) & \textit{if} \quad u > 0 \\ (\theta - 1)\,u & \textit{if} \quad u \leq 0 \end{cases}$$

On the objective function

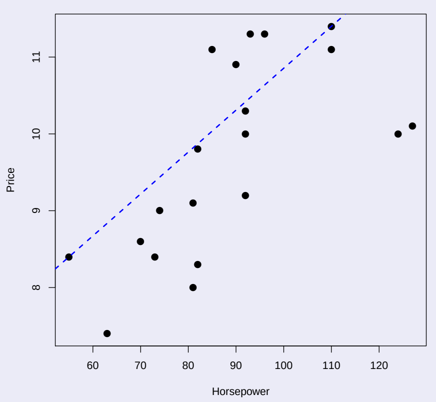
θ=0.25



- 75% of points above the QR line and 25% below
- unbalanced weighting system: 0.75 (0.25) for sum of negative (positive) deviations
- *m*=2 points lies exactly on the line (*m*=number of model parameters)

On the objective function

θ=0.75



- 25% of points above the QR line and 75% below
- unbalanced weighting system: 0.25 (0.75) for sum of negative (positive) deviations
- *m*=2 points lies exactly on the line (*m*=number of model parameters)

The linear programming formulation of the QR problem

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

Methods for solving the linear programming problem

- The simplex method (Dantzig, 1947) is the widespread solution for the linear programming problem
- It is an iterative process, starting from a solution that satisfies the imposed constraints and looking for new and better solution
- The process iterates until a solution that cannot be further improved is reached, moving along the edges of the simplex corresponding to the feasible set
- For the QR problem, the efficient version of the simplex algorithm, proposed by Barrodale and Roberts (1974) and adapted by Koenker e D'Orey (1987) to compute conditional quantiles, is typically used with a moderate size problem
- The simplex method is the default option in most of the QR software
- A completely different method approaches the solution from the interior of the feasible set rather than on its boundary, that is starting in the zone where all the inequalites are strictly satisfied
- Such methods, called interior–point methods, have their roots in the seminal paper of Karmakar (1984) and are usually superior on very large problems
- The QR solution using interior–point methods has been proposed by Portnoy e Koenker (1997)
- A heuristic approach (finite smoothing algorithm) has been proposed by Chen (2004, 2007): it is faster and more accurate in the presence of a large number of covariates

Part 1: Quantile Regression

Inference

Main approaches to inference in QR

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

Main approaches to inference in QR

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

Asymptotic theory

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

"*under mild regularity conditions*"

$$\Downarrow$$

Asymptotic distribution of the estimator:

1. case of i.i.d. errors

$$\sqrt{n}\left[\hat{\beta}(\theta) - \beta(\theta)\right] \to N\left(0, \varpi^2(\theta)\,\mathbf{J}^{-1}\right)$$

2. case of i.ni.d. errors

$$\sqrt{n}\left[\hat{\beta}(\theta) - \beta(\theta)\right] \to N\left(0, \theta(1-\theta)\,\mathbf{H}(\theta)^{-1}\,\mathbf{J}\mathbf{H}(\theta)^{-1}\right)$$

The error distribution affects the variance–covariance matrix of the QR estimator

Main approaches to inference in QR

Asymptotic theory

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

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

Bootstrap approach

- useful when the assumptions for the asymptotic procedure do not hold
- easy to compute standards errors
- flexible to obtain standard error and confidence interval for any estimates and combinations of estimates

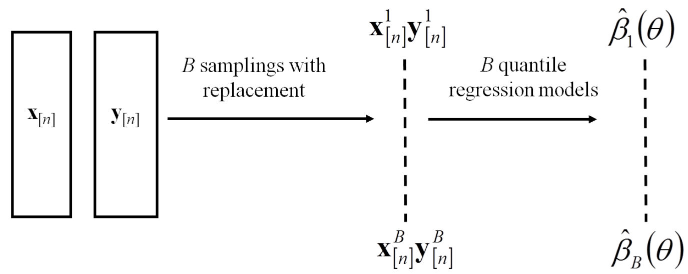
Resampling methods in QR

- *xy-pair or design matrix bootstrap method* (Kocherginsky, 2003)
- method based on pivotal estimation functions (Parzen, 1979)
- markov chain marginal bootstrap (He and Hu, 2002) (Kocherginsky, 2003) (Kocherginsky et al. 2005)

xy-pair method: a single quantile θ

Simple quantile regression model

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



Bootstrap estimate: $\overline{\hat{\beta}}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b(\theta)$

Bootstrap standard error: $se\left(\hat{\beta}_j(\theta_q)\right)$

Part 1: Quantile Regression

Equivariance properties

Equivariance properties

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

Equivariance to monotone transformations

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

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

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

- The quantiles of the transformed **y** variable are the transformed quantiles of the original ones
- appropriate selection of $h(.)$ corrects different kinds of skewness
- 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)

Part 1: Quantile Regression

Assessment

QR assessment

- Quantile regression models are estimated minimizing the absolute values of weighted residuals, as opposed to minimizing the sum of squared errors in OLS
- The R2 is not an applicable goodness-of-fit measure
- Methods available for evaluating goodness-of-fit in quantile regression allow to compare model fit among nested model but they are not comparable to standard coefficients of determination

Koenker R. and Jose A.F. Machado. Goodness of Fit and Related Inference Processes for Quantile Regression J. of Am Stat. Assoc, (1999), 94, 1296-1310

QR assessment

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

Residual absolute sum of weighted differences:

$$RASW_\theta = \sum_{y_i \geq \beta_0(\theta)+\beta_1(\theta)x_i} \theta \left| y_i - \beta_0(\theta) - \beta_1(\theta)x_i \right| + \sum_{y_i < \beta_0(\theta)+\beta_1(\theta)x_i} (1-\theta) \left| y_i - \beta_0(\theta) - \beta_1(\theta)x_i \right|$$

Model: $Q_\theta(\hat{\mathbf{y}}) = \hat{\beta}_0(\theta)$

Total absolute sum of weighted differences:

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

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

Part 1: Quantile Regression

An empirical analysis

An empirical analysis

The aim of the analysis

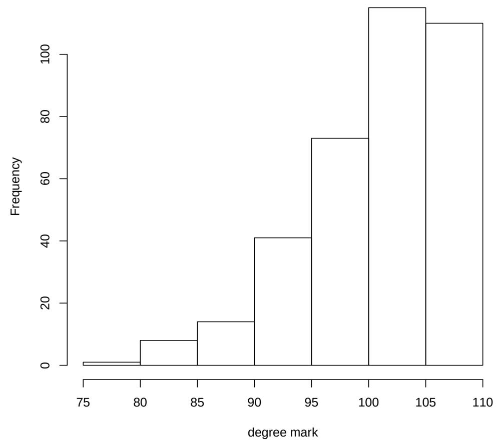
Evaluate if and how
the **student features**
(socio-demographic and University experience attributes)
affect the **outcome** of the University career (degree mark) in case of unobserved group heterogeneity

The dataset

The evaluation of University educational processes

- random sample of 362 students graduated at University of Macerata (Italy)
- dependent variable: degree mark (110 scores excluded)
- 7 regressors related to the student profile:
  - gender
  - place of residence during University (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 get a degree
  - diploma mark

The dataset



Distribution of the dependent variable

OLS and QR coefficients

	OLS	θ=0.10	θ=0.25	θ=0.50	θ=0.75	θ=0.90
(Intercept)	**101.78**	**100.12**	**101.08**	**102.19**	**103.60**	**106.45**
Gender = Male	**-3.42**	-1.94	**-3.92**	**-4.12**	**-2.60**	**-1.38**
Place of residence = Marche region	0.95	0.89	**1.69**	1.33	**1.05**	0.17
Place of Residence = outside Marche	**-2.51**	**-8.19**	-2.50	-2.04	-0.95	-0.79
Courses attendance = regular	**1.87**	2.52	0.92	**2.34**	**1.25**	**1.25**
Working student = yes	-0.20	0.62	0.42	-0.21	-0.60	-0.31
Numbers of years to get a degree	**-0.82**	**-1.27**	**-1.42**	**-0.88**	-0.35	-0.17
Diploma mark	**0.06**	0.01	0.08	0.07	**0.05**	0.02

Outline

Heterogeneity

Part 1: Quantile Regression

- Basic insights
- Estimation
- Inference
- Properties
- Assessment

Part 2: My recent research on handling heterogeneity

- Unsupervised approach
- Supervised approach
- Quantile Composite-based Path Model

Outline

Heterogeneity

Part 1: Quantile Regression

- Basic insights
- Estimation
- Inference
- Properties
- Assessment

Part 2: My recent research on handling heterogeneity

- Unsupervised approach
- Supervised approach
- Quantile Composite-based Path Model

Handling heterogeneity among units

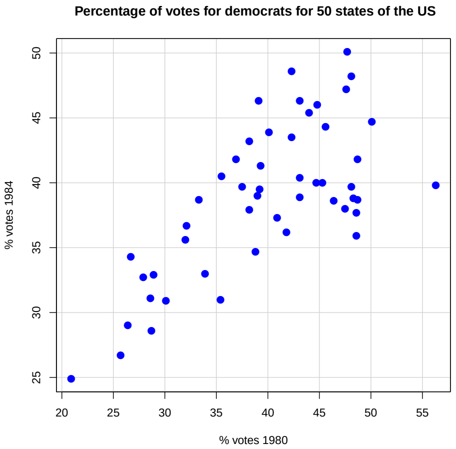
Identification of group effects in a regression model

- Unsupervised approach
- Supervised approach

CLUSTERING & MODELING: Identifying a typology in a dependence model

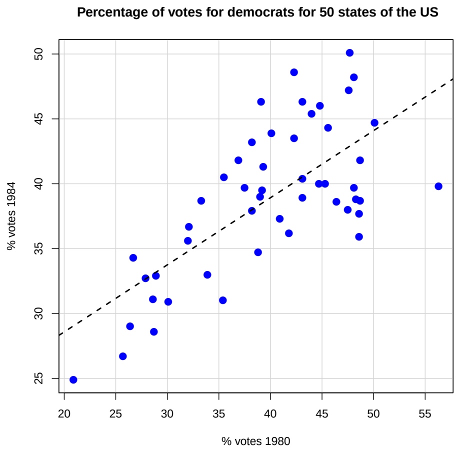
- Identifying groups of units characterized by similar dependence structures
- Discovering the best model for each group
- Testing differences among groups

A simple example

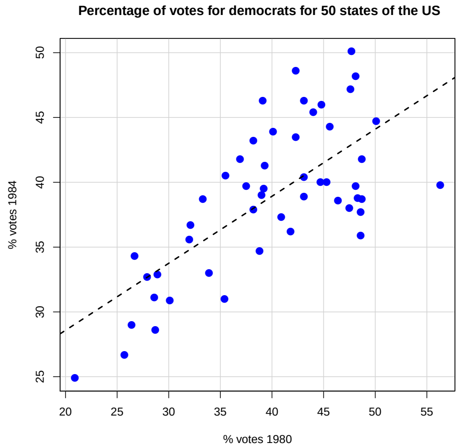


http://rcarbonneau.com/ClusterwiseRegressionDatasets.htm

A simple example



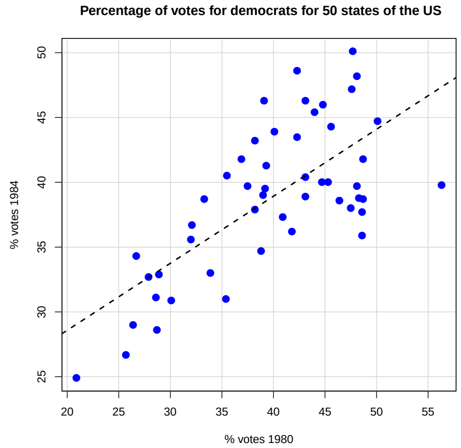
A simple example



Research questions?

- How to identify unobserved heterogeneity?
- How to partition the units according to the dependence relationship?
- How many groups?

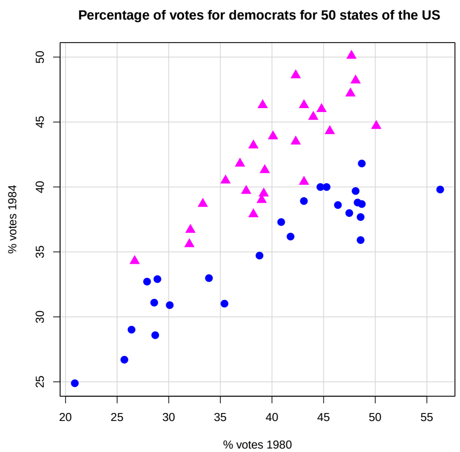
A a simple example



Research questions?

- How to identify unobserved heterogeneity?
- How to partition the units according to the dependence relationship?
- How many groups?

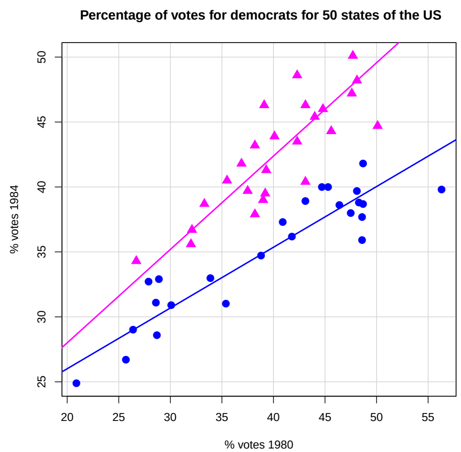
A simple example



Research questions?

- How to identify unobserved heterogeneity?
- How to partition the units according to the dependence relationship?
- How many groups?
- What is the best model for each group?

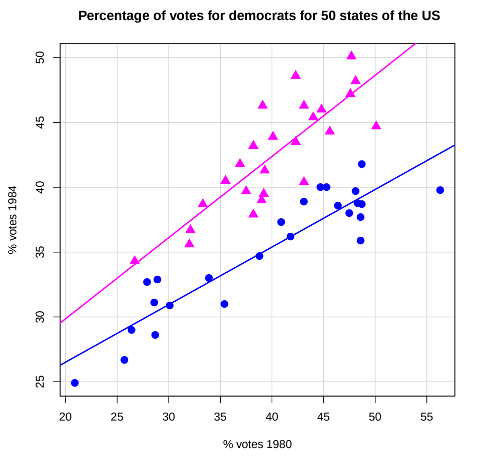
A simple example



Research questions?

- How to identify unobserved heterogeneity?
- How to partition the units according to the dependence relationship?
- How many groups?
- What is the best model for each group?

A simple example



Research questions?

- How to identify unobserved heterogeneity?
- How to partition the units according to the dependence relationship?
- How many groups?
- What is the best model for each group?

The main steps

1. Identification of the global dependence structure
2. Identification of the best model for each unit
3. Clustering units
4. Modeling groups
5. Testing differences among groups

Basic notation

The data structure

- n units
- p regressors
- 1 quantitative or ordinal dependent variable

$\mathbf{y}_{[n]}$ $\mathbf{X}_{[n \times p]}$

Basic notation

The data structure

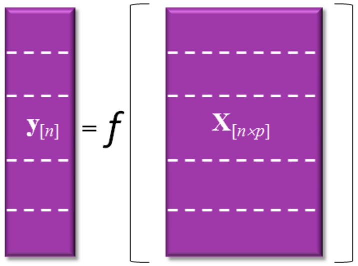
- n units
- p regressors
- 1 quantitative or ordinal dependent variable

$$\mathbf{y}_{[n]} = f\left[\mathbf{X}_{[n \times p]}\right]$$

Basic notation

The data structure

- n units
- p regressors
- *1* quantitative or ordinal dependent variable



G unknown groups

A working example: 2 groups

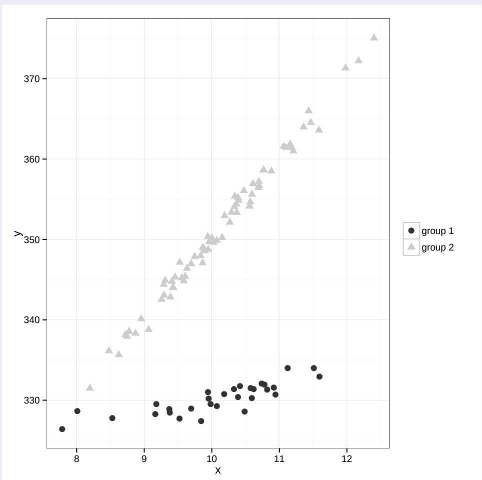
Structure of the two groups

	group 1	group 2
sample size	$n_1 = 30$	$n_2 = 70$
regressor	$\mathbf{x}_1 \sim N(10;1)$	$\mathbf{x}_2 \sim N(10;1)$
error	$\mathbf{e}_1 \sim N(0;1)$	$\mathbf{e}_2 \sim N(0;1)$
response variable	$\mathbf{y}_1$=310+2$\mathbf{x}_1$+$\mathbf{e}$	$\mathbf{y}_2$=250+10$\mathbf{x}_2$+$\mathbf{e}$

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \tag{2}$$

A working example

Structure of the two groups



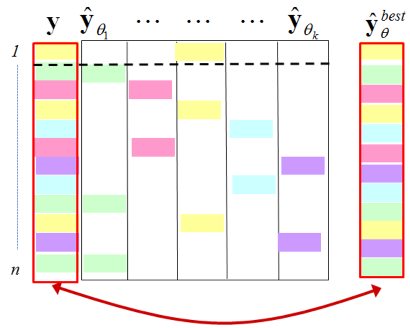
The proposed approach

1.Identification of the global dependence structure

$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta) \quad \theta = 1,\dots,k$

2.Identification of the best model for each unit

- estimated values
  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
- best model identification
  θ_i^{best} : $\underset{\theta=1,\dots,k}{\text{argmin}} \; |\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)|$
- best estimates identification
  $\hat{\mathbf{y}}_\theta^{best}$



A working example: 2 groups

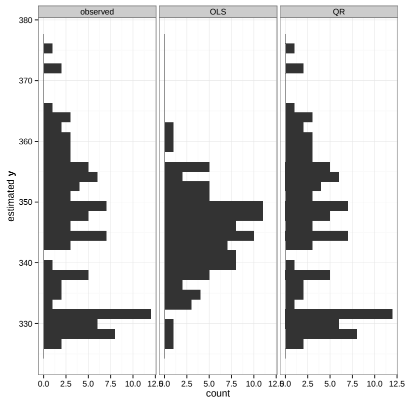
1. Global estimation

$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$

2. Identification of the best model for each unit

1. estimated values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. best model identification $\theta_i : \underset{\theta=1,\ldots,k}{\text{argmin}} \; |y_i - \hat{y}_i(\theta)|$
3. best estimates identification $\hat{\mathbf{y}}_\theta^{best}$

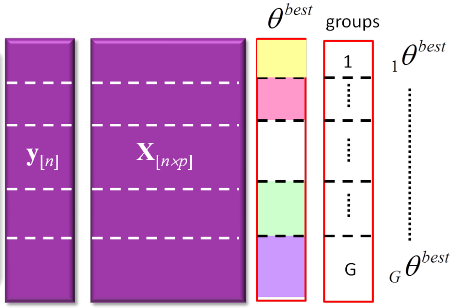
Distribution of the dependent variable: observed (left panel), LS estimated (middle panel), best QR estimated (right panel)



The proposed approach

3. Clustering units

- finding the best partition of the θ^{best} vector
- identification of the group reference quantile ${}_g\theta^{best}$, for $g = 1, G$



3. Clustering units

Finding the best partition of the θ^{best} vector

- θ^{best} is partitioned into D groups (e.g. according to the deciles)
- identification of a reference quantile for each of the D groups:

$$_d\bar{\theta}^{best} = \frac{\sum_{i=1}^{n_d} \theta_i^{best}}{n_d}$$

$(d = 1, \ldots, D)$

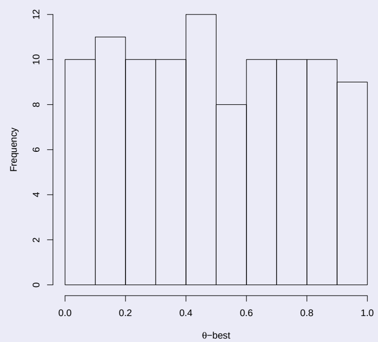
- estimate D quantile regression models with $\theta = \left[{}_1\bar{\theta}^{best}, \ldots, {}_D\bar{\theta}^{best}\right]$

A working example: 2 groups

3. Clustering units

Finding the best partition of the θ^{best} vector: a solution

- θ^{best} is partitioned according to its deciles $(d = 1, \ldots, D)$

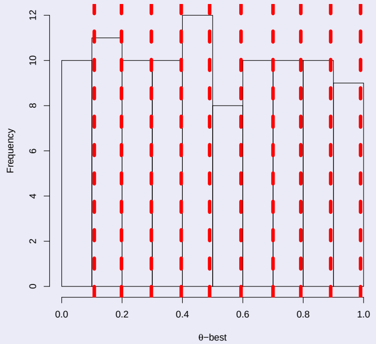


A working example: 2 groups

3. Clustering units

Finding the best partition of the θ^{best} vector

- θ^{best} is partitioned according to its deciles ($d = 1, \ldots, D$)



A working example: 2 groups

3. Clustering units

Finding the best partition of the θ^{best} vector

- identification of a reference quantile for each of the D groups:

quantile	value	${}_d\bar{\theta}^{best}$
0.1	0.108	0.046
0.2	0.198	0.148
0.3	0.297	0.246
0.4	0.396	0.345
0.5	0.490	0.435
0.6	0.594	0.545
0.7	0.700	0.642
0.8	0.792	0.750
0.9	0.891	0.845

- estimate D quantile regression models

3. Clustering units

Finding the best partition of the θ^{best} vector

- test whether the slopes of pairs of consecutive models are identical

  Joint Test of Equality of Slopes

  Koenker R.W. and Basset G. 1982 Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**(1)

- group units if their reference quantiles do not provide significantly different coefficients

- identification of the group reference quantile ${}_g\theta^{best}$, for $g = 1, G$

Heteroschedasticity test

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

$H_0 : \beta_1(\theta_i) = \beta_1(\theta_j)$

Test Statistic:

$$T = \frac{\left[\hat{\beta}_1(\theta_i) - \hat{\beta}_1(\theta_j)\right]^2}{var\left[\hat{\beta}_1(\theta_i) - \hat{\beta}_1(\theta_j)\right]} \sim \chi^2_{1gdl} \qquad (3)$$

where $var\left[\hat{\beta}_1(\theta_i) - \hat{\beta}_1(\theta_j)\right] =$
$var\left[\hat{\beta}_1(\theta_i)\right] + var\left[\hat{\beta}_1(\theta_j)\right] - 2cov\left[\hat{\beta}_1(\theta_i)\,\hat{\beta}_1(\theta_j)\right]$

A possible solution to estimate $var\left[\hat{\beta}_1(\theta_i) - \hat{\beta}_1(\theta_j)\right]$: bootstrap

A working example: 2 groups

3. Clustering units

Finding the best partition of the θ^{best} vector

- sequentially test if the slope coefficients of the models are identical

quantile	value	$_d\bar{\theta}^{best}$	p-value
0.1	0.108	0.046	0.853
0.2	0.198	0.148	0.872
0.3	0.297	0.246	0.000
0.4	0.396	0.345	0.758
0.5	0.490	0.435	0.975
0.6	0.594	0.545	0.489
0.7	0.700	0.642	0.152
0.8	0.792	0.750	0.660
0.9	0.891	0.845	0.912

A working example: 2 groups

3. Clustering units

Finding the best partition of the θ^{best} vector

- group units if their reference quantiles provide not significantly different coefficients

quantile	value	$_d\bar{\theta}^{best}$	p-value	group	n_g
0.1	0.108	0.046	0.853	1	30
0.2	0.198	0.148	0.872		
0.3	0.297	0.246	0.000		
0.4	0.396	0.345	0.758	2	70
0.5	0.490	0.435	0.975		
0.6	0.594	0.545	0.489		
0.7	0.700	0.642	0.152		
0.8	0.792	0.750	0.660		
0.9	0.891	0.845	0.912		

A working example: 2 groups

3. Clustering units

Finding the best partition of the θ^{best} vector

- identification of the group reference quantile

quantile	value	$_d\bar{\theta}^{best}$	p-value	group	n_g	$_g\theta^{best}$
0.1	0.108	0.046	0.853	1	30	0.147
0.2	0.198	0.148	0.872			
0.3	0.297	0.246	0.000			
0.4	0.396	0.345	0.758	2	70	0.649
0.5	0.490	0.435	0.975			
0.6	0.594	0.545	0.489			
0.7	0.700	0.642	0.152			
0.8	0.792	0.750	0.660			
0.9	0.891	0.845	0.912			

The proposed approach

4. Modeling groups

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(_g\theta^{best})$$

5. Testing differences among groups

- Testing if all the slope coefficients of the groups are identical
- Separate testing on each slope coefficient

Koenker R.W. and Basset G. 1982 Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**(1)

A working example: 2 groups

4. Modeling groups

	$\theta = 0.145$ group 1	$\theta = 0.640$ group 2
intercept	313.11	248.19
x	1.71	10.19
original model	$\mathbf{y}_1$=310+2$\mathbf{x}_1$+$\mathbf{e}$	$\mathbf{y}_2$=250+10$\mathbf{x}_2$+$\mathbf{e}$

Percentage of Correct classification (%CC)=100%

An empirical analysis

The aim of the analysis

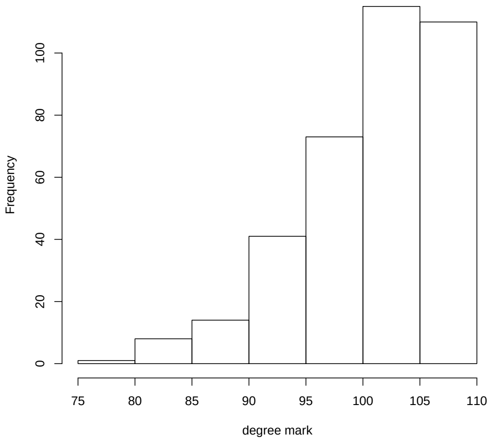
Evaluate if and how
the **student features**
(socio-demographic and University experience attributes)
affect the **outcome** of the University career (degree mark) in case of unobserved group heterogeneity

The dataset

The evaluation of University educational processes

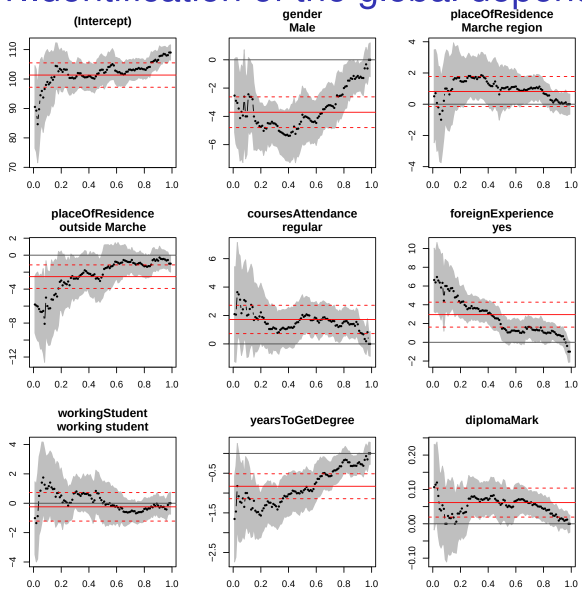
- random sample of 362 students graduated at University of Macerata (Italy)
- dependent variable: degree mark (110 scores excluded)
- 7 regressors related to the student profile:
  - gender
  - place of residence during University (Macerata and its province, Marche region, outside Marche)
  - course attendance (no attendance, regular)
  - foreign experience (yes, no)
  - working condition (full time student, working student)
  - number of years to get a degree
  - diploma mark

The dataset



Distribution of the dependent variable

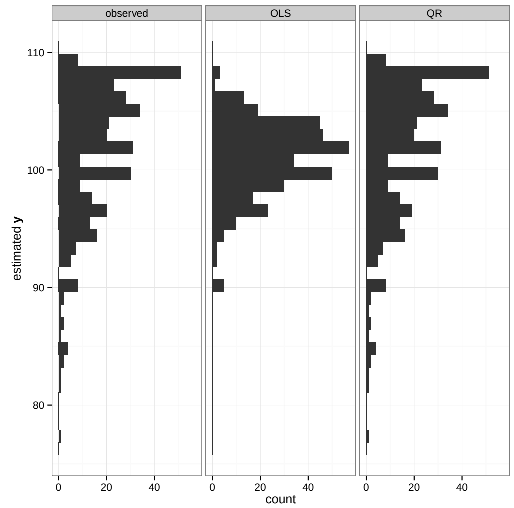
1.Identification of the global dependence structure



LS and QR coefficients

Step 1:
$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta) \quad \theta = 1, \ldots, k$

Step 2: Identification of the best model for each unit



Distribution of the:

- dependent variable (*left panel*)
- LS estimated dependent variable (*middle panel*)
- best QR estimated dependent variable (*right panel*)

Step 2:

- estimated values: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
- best model identification $\theta_i^{best}: \underset{\theta=1,\ldots,k}{\text{argmin}} \ |\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)|$
- best estimates identification: $\hat{\mathbf{y}}_\theta^{best}$

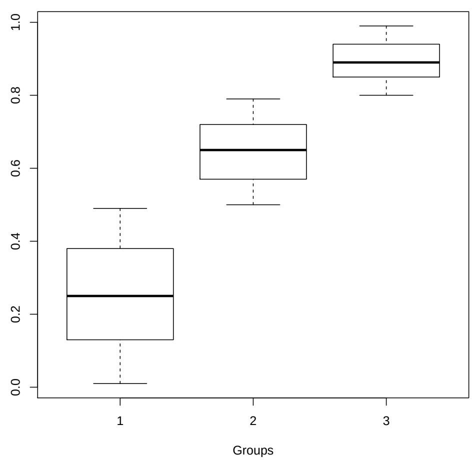
Step 3: Clustering units

quantile	value	$_d\bar{\theta}^{best}$	p-value	group	n_g	$_g\theta^{best}$
0.1	0.090	0.036	0.412	1	182	0.246
0.2	0.190	0.145	0.170			
0.3	0.293	0.250	0.842			
0.4	0.400	0.341	0.631			
0.5	0.490	0.444	0.000			
0.6	0.596	0.547	0.322	2	109	0.650
0.7	0.690	0.636	0.168			
0.8	0.790	0.747	0.008			
0.9	0.889	0.844	0.298	3	71	0.896

Step 3:

- partitioning of θ^{best}
- identification of the group reference quantile $_g\theta^{best}$, for $g = 1, G$

Step 3: Clustering units

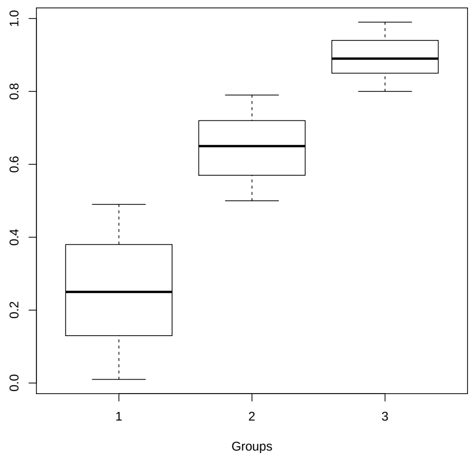


Distribution of the “best” quantiles in the groups

Step 3:

- partitioning of θ^{best}
- identification of the group reference quantile $_g\theta^{best}$, for $g = 1, G$

Step 3: Clustering units



Reference 'best" quantile for each group:

Mean value of the "best" quantiles assigned to units belonging to the g^{th} group

- θ_1^{best}=0.246
- θ_2^{best}=0.649
- θ_3^{best}=0.896

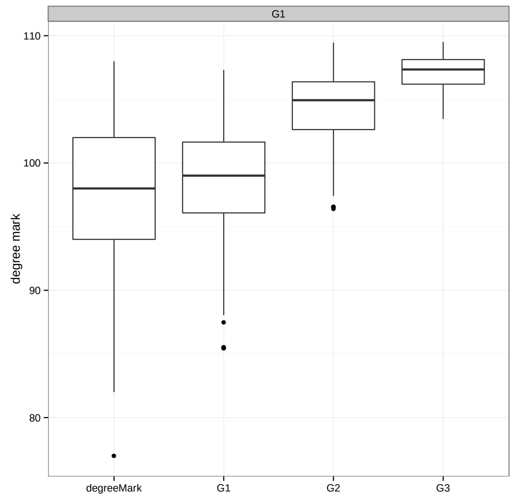
Step 4: $Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}({}_g\theta^{best})$

Step 4: Modeling groups

QR coefficients with group effects

Variable	OLS	G1 $\theta = 0.246$	G2 $\theta = 0.649$	G3 $\theta = 0.896$
Intercept	**101.35**	**102.74**	**101.43**	**106.43**
gender (Male)	**-3.71**	**-5.04**	**-3.61**	**-1.14**
place of residence (Marche region)	0.81	**1.64**	**0.88**	0.25
place of residence (outside Marche)	**-2.53**	**-3.60**	-0.63	-0.64
courses attendance (regular)	**1.72**	0.99	**1.83**	**1.40**
foreign experience (yes)	**2.95**	**3.38**	1.09	0.76
working student	-0.24	-0.17	-0.49	-0.14
years to get a degree	**-0.83**	**-1.22**	**-0.52**	**-0.25**
diploma mark	**0.06**	0.04	**0.07**	0.02

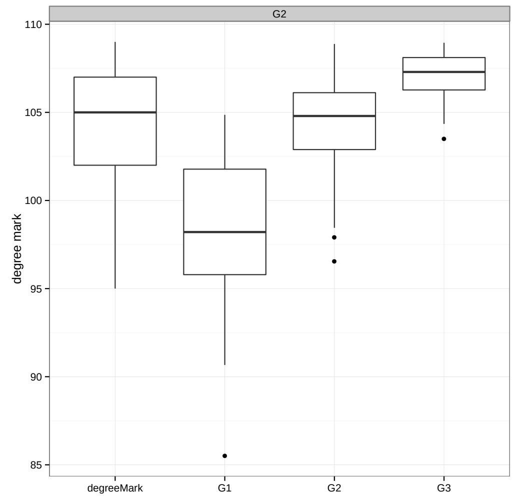
Step 4: Modeling groups



Group 1

Observed response distribution compared with the estimated distributions using the reference quantile of G1

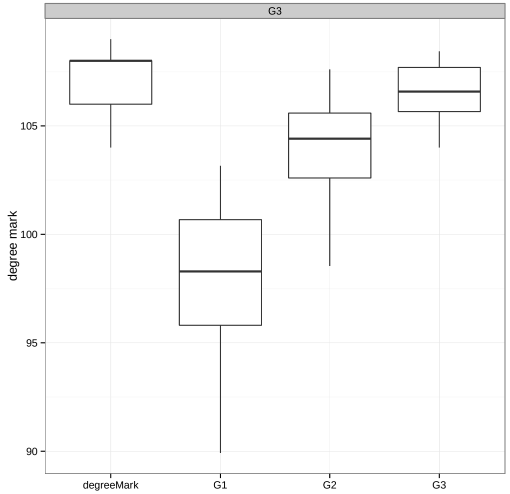
Step 4: Modeling groups



Group 2

Observed response distribution compared with the estimated distributions using the reference quantile of G2

Step 4: Modeling groups



Group 3

Observed response distribution compared with the estimated distributions using the reference quantile of G3

Step 5: Testing differences among groups

Testing if all the slope coefficients of the groups are identical

p-values

	G1	G2	G3	G1;G2;G3
G1		0.001021	0.000000	
G2			0.000329	
G3				0.000000

Separate testing on each slope coefficient

	g1 vs g2	g2 vs g3	g1 vs g3
gender (Male)	0.114	0.003	0.000
place of residence (Marche region)	0.202	0.131	0.024
place of residence (outside Marche)	0.051	0.990	0.081
courses attendance (regular)	0.253	0.484	0.599
foreign experience (yes)	0.005	0.646	0.000
working student	0.609	0.436	0.969
years to get a degree	0.008	0.115	0.000
diploma mark	0.341	0.006	0.549

Recap & Pros

Clustering units taking into account the dependence structure

- Estimation of the group dependence structure using the whole sample
- Impact of the regressors on the entire conditional distribution
- Clarity of the final results
- Availability of classical inferential procedures to test differences among groups
- Number of groups defined by the procedure
- Exact solution method

Further developments

- Explore alternatives to partition th θ_{best} vector
- Introduce cluster validation statistics
- Simulation study
- Comparison with competitive methods

A simulation study

Exploring the robustness of the method with respect to:

1. the degree and type of overlapping among the groups;
2. the cardinality of each group (equal or unbalanced);
3. the sample size.

case of one regressor and two groups

Generation of a set of scenarios:

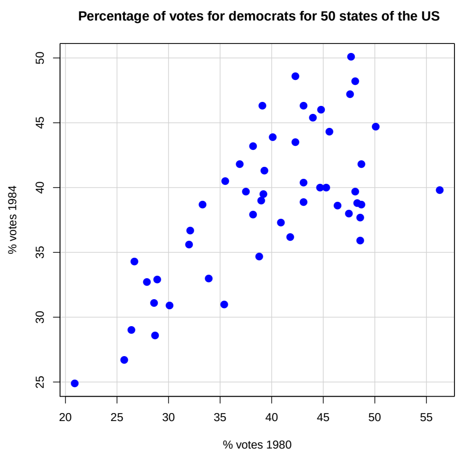
- Case 1 : parallel group structures;
- Case 2 : group structures crossing outside the considered range of the regressor;
- Case 3 : group structures crossing inside the considered range of the regressor.

Comparison with competitive methods

Clusterwise linear regression

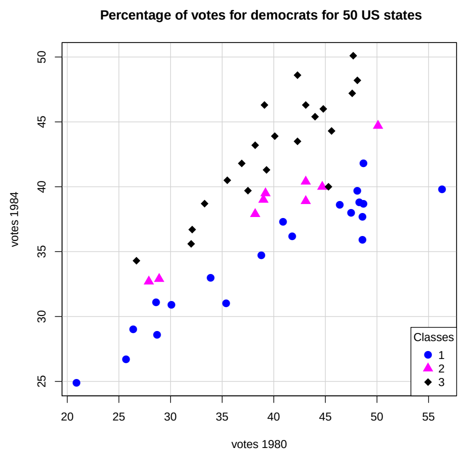
- It is a useful technique when heterogeneity is present in the data
- It identifies both the partition of the data and the relevant regression models, one for each cluster.
- It estimates simultaneously the classes and the parameters of the models which are considered different on each class
- Number of classes a-priori defined
- Not exact solution method
- Performance is sensitive to the initial partition and outliers
- Overlapping among groups

A comparison with the 'votes' dataset



http://rcarbonneau.com/ClusterwiseRegressionDatasets.htm

A comparison with the 'votes' dataset

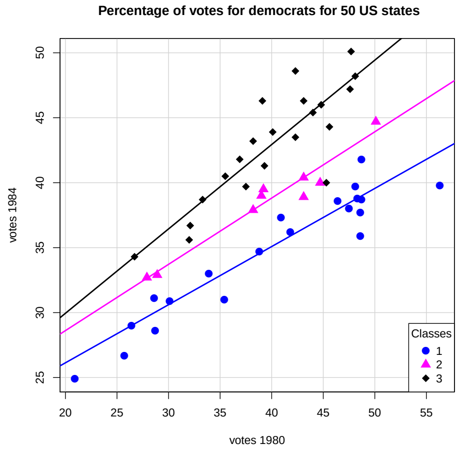


The proposed approach

Best partition: 3 groups

- θ_1^{best}=0.18
- θ_2^{best}=0.49
- θ_3^{best}=0.79

A comparison with the ‘votes’ dataset

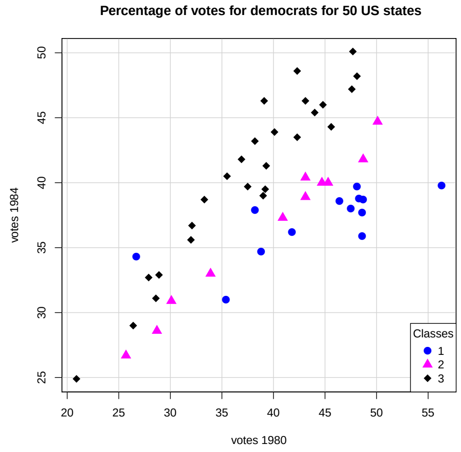


The proposed approach

Best partition: 3 groups

- θ_1^{best}=0.18
- θ_2^{best}=0.49
- θ_3^{best}=0.79

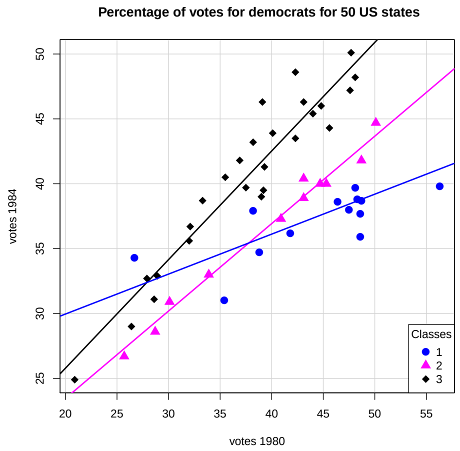
A comparison with the ‘votes’ dataset



Clusterwise linear regression

- a-priori definition of the number of classes
- No exact solution method
- Performance is sensitive to the initial partition and outliers
- Overlapping among groups

A comparison with the ‘votes’ dataset



Clusterwise linear regression

- a-priori definition of the number of classes
- No exact solution method
- Performance is sensitive to the initial partition and outliers
- Overlapping among groups

Comparison with alternative methods

Research questions to be explored

- How to compare results?
- What are other alternative methods?
- ....
- ....

Handling heterogeneity among units

Identification of group effects in a regression model

- Unsupervised approach
- Supervised approach

Comparison with alternative methods

- Estimation of different models for each group
- Introduction of a dummy variable
- Multilevel modeling

Outline

Heterogeneity

Part 1: Quantile Regression

- Basic insights
- Estimation
- Inference
- Properties
- Assessment

Part 2: My recent research on handling heterogeneity

- Unsupervised approach
- Supervised approach
- Quantile Composite-based Path Model

Concluding remarks: motivation

Motivation (Mosteller and Tukey, 1977)

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of **X**'s.
We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set.
Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

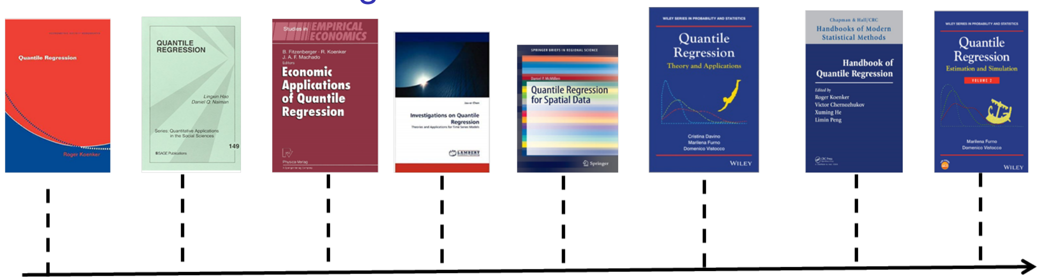
Concluding remarks: motivation

QR is capable of providing a more complete, more nuanced view of heterogeneous covariate effects (Koenker et al., 2017)

Caution

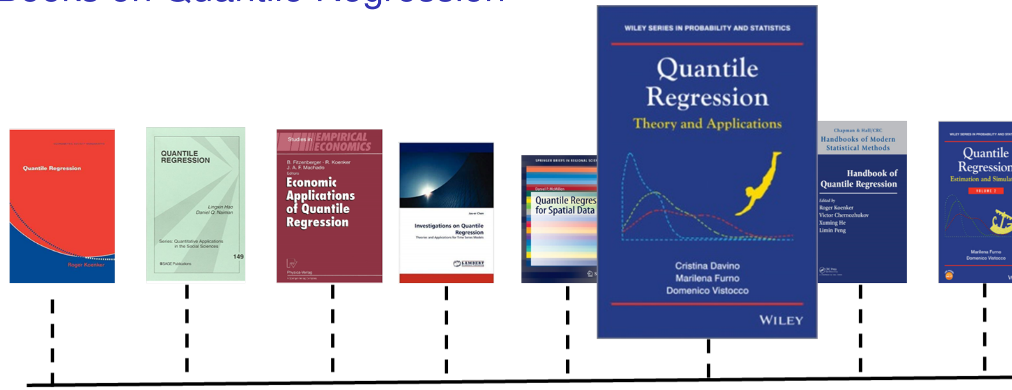
QR offers information on the whole conditional distribution of the response variable, allowing us to discern effects that would otherwise be judged equivalent using only conditional expectation.
Nonetheless, the QR ability to statistically detect more effects can not be considered a panacea for investigating relationships between variables: in fact, the improved ability to detect a multitude of effects forces the investigator to clearly articulate what is important to the process being studied and why.

Books on Quantile Regression



2005	2007	2010	2010	2012	2013	2017	2018
R. Koenker *Quantile Regression.* Econometric Society Monographs	L. Hao, D. Q. Naiman *Quantile Regression.* Series Quantitative Applications in the Social Sciences - Sage Publications		Jau-er Chen Investigations on Quantile Regression. Theories and Application for Time Series Models LAP Lambert Academic Publishing	Daniel P. (2012) *Quantile Regression for Spatial Data .* SpringerBriefs in Regional Science	C. Davino, M. Furno, D. Vistocco *Quantile Regression: Theory and Applications,* Wiley Series in probability and Statistics	R. Koenker, V. Chernozhukov, X. He, L. Peng (eds) *Handbook of Quantile Regression* Chapman and Hall/CRC	M. Furno, D. Vistocco *Quantile Regression: Estimation and Simulation,* Volume 2, Wiley Series in probability and Statistics

Books on Quantile Regression



2005	2007	2010	2010	2012	2013	2017	2018
R. Koenker *Quantile Regression.* Econometric Society Monographs	L. Hao, D. Q. Naiman *Quantile Regression.* Series Quantitative Applications in the Social Sciences - Sage		Jau-er Chen Investigations on Quantile Regression. Theories and Application for Time Series Models LAP Lambert Publishing	Regional		Hall/CRC	M. Furno, Vistocco *Quantile Regressio Estimatio and Simulatio* Volume 2 Wiley Ser in probab and Statis

C. Davino, M. Furno, D. Vistocco
Quantile Regression:
Theory and Applications
Wiley Series in probability and Statistics

Main references

- Davino C., Vistocco D. (2018) Handling heterogeneity among units in quantile regression. Investigating the impact of students' features on University outcome, *Statistics & Its Interface*, Vol. 11, pp. 541-556.
- Davino C., Romano R., Vistocco D. (2018) Modelling drivers of consumer liking handling consumer and product effects, *Italian Journal of Applied Statistics* (in press).
- Davino C., NaesT., Romano R., Vistocco D. (2018) Modeling preferences: beyond the average effects, In Capecchi S., Di Iorio F., Simone R. (eds) *ASMOD 2018 : Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference* pp. 93-100, FedOAPress, Napoli.
- Davino C., Dolce P., Taralli S. (2017) Quantile Composite-based Model: a Recent Advance in PLS-PM. A Preliminary Approach to Handle Heterogeneity in the Measurement of Equitable and Sustainable Well-Being. In H. Latan, R. Noonan (eds) *Partial Least Squares Structural Equation Modeling - Basic Concepts, Methodological Issues and Applications*, Springer.
- Davino C., Dolce P., Esposito Vinzi V., Taralli S. (2016) A Quantile Composite-Indicator Approach for the Measurement of Equitable and Sustainable Well-Being: A Case Study of the Italian Provinces. *Social Indicators Research*, vol. xx, p. 1-318.
- Davino C., Esposito Vinzi V. (2016) Quantile Composite-based Path Modelling, *Advances in Data Analysis and Classification. Theory, Methods, and Applications in Data Science*, vol. 10, pp. 491-520.

Main references

- Davino C., Esposito Vinzi V., Dolce P. (2016) Assessment and Validation in Quantile Composite-Based Path Modeling. In: (a cura di): H. Abdi, V. Esposito Vinzi, G. Russolillo, G. Saporta, L. Trinchera, *The Multiple Facets of Partial Least Squares and Related Methods*. p. 169-185, Springer International Publishing.
- Davino C., Vistocco D. (2015) Quantile Regression for Clustering and Modeling Data. In I. Morlini, T. Minerva, M. Vichi (eds) *Advances in Statistical Models for Data Analysis*: Studies in Classification, Data Analysis, and Knowledge Organization. p. 85–96, Springer, Heidelberg.
- Davino C., Furno M., Vistocco D. (2013) *Quantile Regression: Theory and Applications*. Wiley.
- Davino C., Vistocco D. (2008) Quantile regression for the evaluation of student satisfaction. *Italian Journal of Applied Statistics* **20**, 179–196.
- Davino C., Romano R., Naes T. (2015) The use of quantile regression in consumer studies. *Food Quality and Preference*, 40, pp. 230–239, Elsevier.
- Davino C., Vistocco D. (2015) Quantile Regression for Clustering and Modeling Data, In I. Morlini, T. Minerva, M. Vichi (eds), Advances in Statistical Models for Data Analysis, pp. 85–95, Springer.
- Davino C., Vistocco D. (2007) The evaluation of University educational processes: a quantile regression approach. *STATISTICA*, n.3, pp. 267-278.

Main references

- Koenker R.W., Basset G. (1978) Regression quantiles. *Econometrica*, Vol. 46, No. 1.
- Koenker R. (2005) *Quantile Regression*. Econometric Society Monographs. Cambridge: Cambridge University Press.
- Koenker R. (2018). quantreg: Quantile Regression. R package version 5.35. https://CRAN.R-project.org/package=quantreg
- Mosteller F., Tukey J. (1977) *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison–Wesley.