Quantile Regression for Group Effect Analysis

Cristina Davino[1] Domenico Vistocco[2]

[1]Dip.to di Studi sullo Sviluppo Economico
Università di Macerata
cdavino@unimc.it

[2]Dip.to di Scienze Economiche
Università di Cassino
vistocco@unicas.it

19th International Conference on Computational Statistics
Paris, 22 – 27 August 2010

all computations and graphics were done in the R language using the packages *quantreg* and *ggplot2*

Outline

Outline

Outline

Outline

Aim of the paper

Identification of group effects in a quantile regression model

1. CONFIRMATIVE APPROACH
2. ROW–PARTITIONED DATA
   - Supervised approach
   - Unsupervised approach

Aim of the paper

Identification of group effects in a quantile regression model

1. CONFIRMATIVE APPROACH
2. ROW–PARTITIONED DATA
   - Supervised approach
   - Unsupervised approach

Aim of the paper

Identification of group effects in a quantile regression model

1. CONFIRMATIVE APPROACH
2. ROW–PARTITIONED DATA
   - Supervised approach
   - Unsupervised approach

Aim of the paper

Identification of group effects in a quantile regression model

1. CONFIRMATIVE APPROACH
2. ROW–PARTITIONED DATA
   - Supervised approach
   - Unsupervised approach

Some solutions for group effect analysis

- Estimation of different models for each group
- Introduction of a dummy variable
- Multilevel modeling (Gelman and Hill, 2007)

Basic notation

The data structure

- n: number of units
- p: number of regressors
- G: number of groups or levels

- $\mathbf{X}_{[n \times p]}$
  - ${}_g x_{ij}$
    $(i$=1,...,n; j=1,...,p; g=1,...$G)$
- $\mathbf{y}_{[n]}$
  - ${}_g y_i$
    $(i$=1,...,n; g=1,...$G)$
- n_g: number of units in group g

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 (Koenker and Basset, 1978) **(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: $(0 < \theta < 1)$

$$\beta_i = \frac{\partial E(\mathbf{y})}{\partial \mathbf{x}_i}$$

$$\beta_i(\theta) = \frac{\partial Q_\theta(\mathbf{y})}{\partial \mathbf{x}_i}$$

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 (Koenker and Basset, 1978) **(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: $(0 < \theta < 1)$

$$\beta_i = \frac{\partial E(\mathbf{y})}{\partial \mathbf{x}_i}$$

$$\beta_i(\theta) = \frac{\partial Q_\theta(\mathbf{y})}{\partial \mathbf{x}_i}$$

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 (Koenker and Basset, 1978) **(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: $(0 < \theta < 1)$

$$\beta_i = \frac{\partial E(\mathbf{y})}{\partial \mathbf{x}_i}$$

$$\beta_i(\theta) = \frac{\partial Q_\theta(\mathbf{y})}{\partial \mathbf{x}_i}$$

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: $(0 < \theta < 1)$

$_g\mathbf{y} = {}_g\mathbf{X}_g\beta +_g \mathbf{e}$

$Q^\theta({}_g\mathbf{y}|{}_g\mathbf{X}) = {}_g\mathbf{X}_g\beta(\theta)$

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: $(0 < \theta < 1)$

$${}_g\mathbf{y} = {}_g\mathbf{X}_g\beta +_g \mathbf{e}$$

$$Q^\theta({}_g\mathbf{y}|{}_g\mathbf{X}) = {}_g\mathbf{X}_g\beta(\theta)$$

The proposed approach

1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\ y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$

The proposed approach

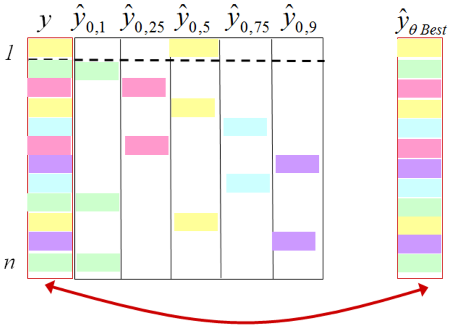
1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\; y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$

The proposed approach

1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\ y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$

The proposed approach

1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\; y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$

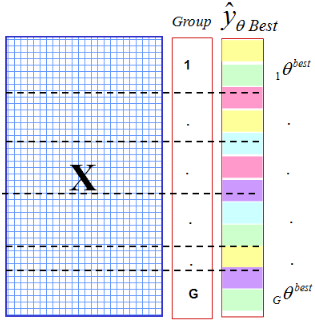


The proposed approach

1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\ y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$

The proposed approach

1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\ y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$



The proposed approach

1. Global estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)$
2. Identification of the best model for each unit
   1. density estimation
      $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta)$
   2. best model identification
      $\theta_i : \underset{\theta=1,\Theta}{\text{argmin}}\ y_i - \hat{y}_i(\theta)$
   3. best density estimation vector
      $\hat{\mathbf{y}}_{\theta}^{best}$
3. Identification of the best model for each group
   ${}_g\theta^{best}$, for $g = 1, G$
4. Partial estimation
   $Q^{\theta}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta)^{best}$

The dataset

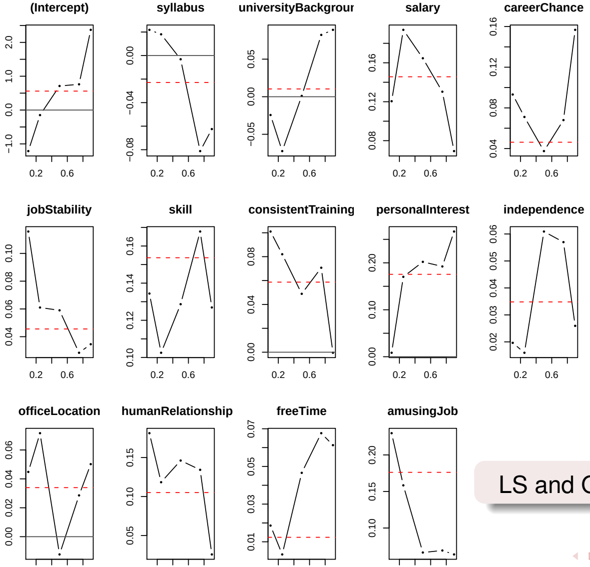
The evaluation of job satisfaction

- n: random sample of 400 students graduated at University of Macerata and in a working condition at the time of the interview
- p: 13 regressors (judgments of the different aspects related to the working experience)

  syllabus, University background, consistent training, career chance, skill, personal interest, free time, salary, office location, job stability, human relationships, amusing job, independence
- dependent variable: overall opinion on the job
- G: 3 groups corresponding to the type of job

  self-employed, private employee, public employee

Step 1: Global estimation



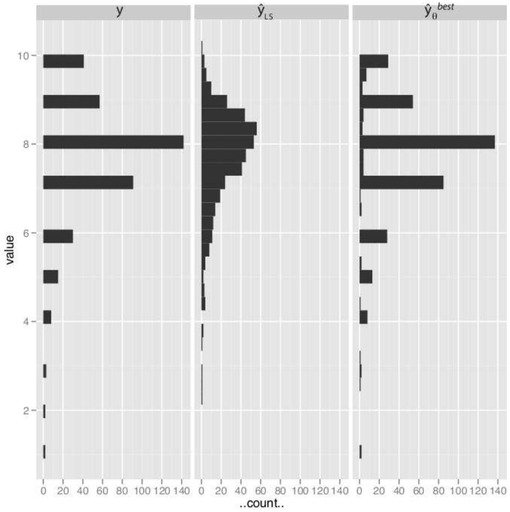
LS and QR coefficients

Step 1: Global estimation

Variable	LS	θ=0.1	θ=0.25	θ=0.5	θ=0.75	θ=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 chance	**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
skill	**0.117**	0.134	0.102	**0.129**	**0.168**	0.127
consistent 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

LS and QR coefficients

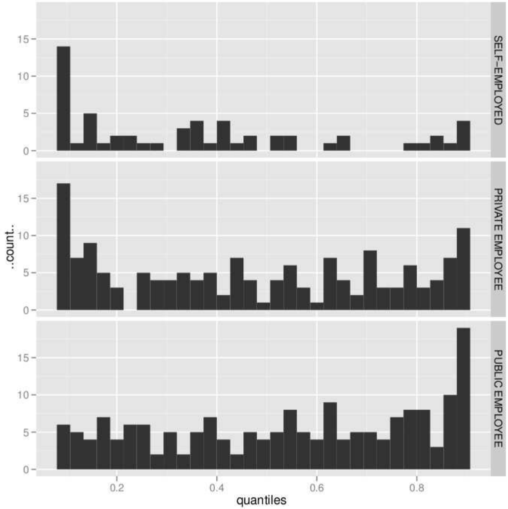
Step 2: Identification of the best model for each unit



Distribution of the:

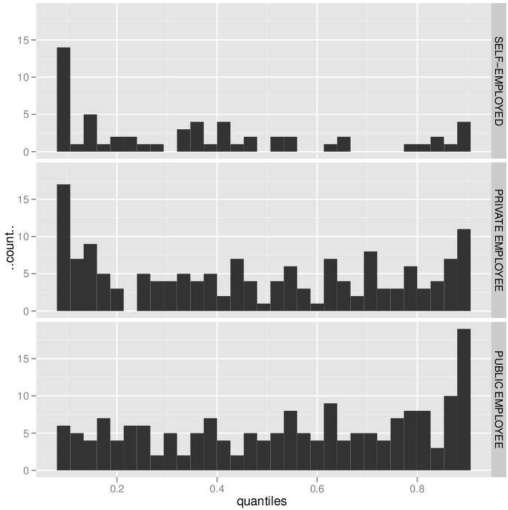
- dependent variable (*left panel*)
- LS estimated dependent variable (*middle panel*)
- best QR estimated dependent variable (*right panel*)

Step 3: Identification of the best model for each group



Distribution of the “best” quantiles assigned to each unit grouped according to the type of job

Step 3: Identification of the best model for each group



"Best" quantiles for each group:

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

- θ_1^{best}=0.371
- θ_2^{best}=0.474
- θ_3^{best}=0.548

Step 4: Partial estimation

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
amusing job	**0.148**	**0.124**	**0.141**

QR coefficients with group effects

Concluding remarks and further issues

The proposed approach

- Group effect analysis
- Impact of the regressors on the entire conditional distribution
- Semi–parametric approach

Further developments

- Robust index for the identification of the "best" quantile
- Statistical significance of the differences among the "best" quantiles
- Time as grouping variable
- Unsupervised approach

Concluding remarks and further issues

The proposed approach

- Group effect analysis
- Impact of the regressors on the entire conditional distribution
- Semi–parametric approach

Further developments

- Robust index for the identification of the “best” quantile
- Statistical significance of the differences among the “best” quantiles
- Time as grouping variable
- Unsupervised approach

Main references

DAVINO, C., VISTOCCO, D. (2007): The evaluation of university educational processes: a quantile regression approach. *STATISTICA*, n.3, pp. 267-278.

DAVINO C., VISTOCCO D (2008): Quantile regression for the evaluation of student satisfaction. STATISTICA APPLICATA, vol. 20; p. 179-196.

EIDE, E. SHOWALTER, M.H. (1998): The effect of school quality on student performance: a quantile regression approach. *Economics Letters 58, 345-350*.

FURNO, M. (2010): Quantile regression analysis of the Italian school system. *Statistical Modelling*, vol. 4, 2010, in press.

GELMAN, A. HILL, J. (2006): *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

HAO, L. NAIMAN, D. Q. (2007): *Quantile Regression*, Series: Quantitative Applications in the Social Sciences, SAGE Publications.

LOCKHEED, M.E. HANUSHECK, E.R.(1994): Concepts of Educational Efficiency and Effectiveness, in Torsten Husén and T. Neville Postlethwaite (ed.), *International Encyclopedia of Education*, second edition, Volume 3 (Oxford: Pergamon, 1994), pp. 1779-1784.

KOENKER, R., BASSET, G.W. (1978): Regression Quantiles, *Econometrica 46, 33-50*.

KOENKER, R. (2005): *Quantile Regression*. Econometric Society Monographs.

KOENKER, R. (2009): *quantreg: Quantile Regression*. R package version 4.44. http://CRAN.R-project.org/package=quantreg.