Pisa, 15 July 2016

Workshop on "Recent Advances in Quantile and M-quantile Regression"

Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: an approach based on quantile regression for counts

Leonardo Grilli *Carla Rampichini*

Dip. di Statistica, Informatica, Applicazioni

Università di Firenze

joint work with *Roberta Varriale* (Istat)

Outline

- Aims
- Data:
  - The pre-enrolment test
  - Administrative records on background characteristics and gained credits
- Quantile regression for counts
  - Introduction
  - Application to gained university credits
- Discussion

Predicting academic performance (so important, so difficult…)

- Predicting students' academic performance is a key step in order to improve the efficiency of university systems
- Universities rely on **info about the high school career**, e.g. type of school and various measures of proficiency
- However, the results at high school are **not fully appropriate** to predict the academic performance:
  - mismatch between competencies evaluated at high school and competencies required for a given degree program
  - heterogeneity in the criteria for awarding marks (variability across types of schools and across geographical regions)
- A (partial) remedy: **pre-enrolment assessment tests** tailored on the needs of each degree program; however, tests have limitations:
  - lack of commonly accepted guidelines
  - shortage of empirical evidence about the predictive ability

Tests vs unstructured interviews

- The results about the predictive ability of pre-enrolment tests are not exciting… what about **unstructured interviews**?
- Apart from the high expense, unstructured interviews are **ineffective** in predicting the students performance:
  - DeVaul R., Jervey F., Chappell J., Caver P., Short B., & O'Keefe S. (1987) Medical school performance of initially rejected students. *Journal of the American Medical Association*, 257, 47-51.
  - Dana J., Dawes R.M., Peterson N.R. (2013) Belief in the Unstructured Interview: The Persistence of an Illusion. *Judgment and Decision Making*, 8(5), pp. 512–520.

"In addition to the vast evidence suggesting that unstructured interviews do not provide incremental validity, we provide direct evidence that **they can harm accuracy**. […] interviewers are likely to feel they are getting useful information from unstructured interviews, even when they are useless. ***Our simple recommendation for those who make screening decisions is not to use them***."

Case study: a pre-enrolment test at the University of Florence

- In a.y. 2008/2009, the School of Economics of the University of Florence introduced a ***compulsory pre-enrolment test*** to evaluate the background of the students
- 40 multiple-choice items covering 3 areas: ***Logic*** (12 items), ***Reading*** (10 items) and ***Mathematics*** (18 items)
  - for each item, one out of 5 alternatives is correct
  - scoring system: 1 if correct, 0 if blank, -0.25 if wrong

- The test has 3 editions (September, November and December)
- ***Candidates with a total score lower than 9 are advised against enrolment***: they could still enrol, but they could take examinations only after 'passing' the test during one of the later editions

`http://www.economia.unifi.it/cmpro-v-p-222.html`

Aim of the research

- For freshmen of the School of Economics – University of Florence, we wish to assess the ***predictive ability*** of a compulsory pre-enrolment test in terms of ***number of gained credits after one year***
- Policy questions:
  - is the pre-enrolment test an effective tool for student self-evaluation?

  - what is its added value with respect to background characteristics already available in administrative records (e.g. type of high school and high school final grade)?
- In statistical terms: conditional on background characteristics, is the test score a good predictor of the number of gained credits?

Dataset for the analysis

- We analyse data on ***690 freshmen*** of the School of Economics in Florence in a.y. 2008/2009, considering the students who took the compulsory pre-enrolment test in September 2008

- The data set is obtained by merging
  - **data collected at the test**
  - **administrative data**

Variables

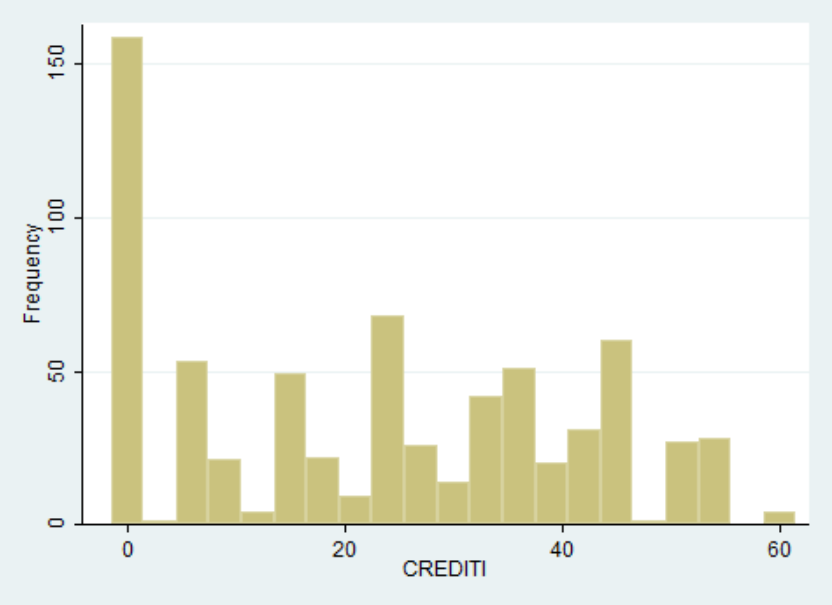
- Pre-test:
  - **Gender**
  - **Far-away resident** (indicator for residence in the provinces of Massa-Carrara and Grosseto or in a province out of Tuscany)
  - **Type of high school** (Scientific, Humanities, Technical, Other)
  - **High school irregular career** (indicator for age at diploma > 19)
  - **High school grade** (from 60 to 100, centered at 80)
- Test: **Total and partial test scores** (Logic, Reading, Mathematics)
- Post-test: **Credits gained during the first year** (from 0 to 60)

How to summarize the test result?

- the three areas (Logic, Reading and Math) have different numbers of items (Math has more weight)
- the three areas may have different predictive power

Thus we do not use the total score, but ***we use the three (standardized) partial scores***

Distribution of gained credits



Gained credits after one year are in the interval [0,60]

Exams have different credits (multiples of 3), usually 6, 9 or 12 → the distribution of gained credits is quite irregular!

- Features of the sample distribution
  - peak at the minimum (23% of freshmen did not gain any credit)
  - the distribution of positive credits is quite irregular, showing peaks at 6, 15, 24, 36 and 45 credits

Flexible modelling approaches

- Due to the features of the sample distribution, we cannot use fully parametric models
- We tried the following approaches:

  - ***Concomitant-variable mixture model*** (not discussed here)
  - ***Hurdle model with quantile regression for counts*** (presented in the following)

- Grilli L., Rampichini C., Varriale R. (2015) Binomial mixture modelling of university credits. *Communications in Statistics - Theory and Methods*. 44(22), pp 4866-4879. http://www.tandfonline.com/eprint/cPGFNfh6saQAAbmkmkdE/full
- Grilli L., Rampichini C. & Varriale R. (2016) Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: an approach based on quantile regression for counts. *Statistical Modelling*. DOI: 10.1177/1471082X15596087

HURDLE MODEL WITH QUANTILE REGRESSION FOR COUNTS

Hurdle (two-part) specification

- We define two sub-models:
  1. for obtaining at least one credit (i.e. a model for the zeroes)
  2. for positive credits
- The **first sub-model** is fitted on the whole population using a logit specification

$$\text{logit}P(Y > 0 \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\alpha}$$

- The **second sub-model** concerns the distribution of credits for those students obtaining at least one credit

$$f(y \mid \mathbf{x}, Y > 0)$$

credits range from 0 to 60 in blocks of 3 → Y = credits/3

no parametric distribution appropriately describes the pattern shown by the credits → to avoid distributional assumptions and account for the discrete nature of credits, we use ***quantile regression for counts***

Quantile regression for counts

- Quantile regression (Koenker 2005) is a methodology to study the relationships between the quantiles of the outcome and a set of covariates, *without any distributional assumption*
  - this approach is very flexible since regression equations at different quantiles are fitted separately, e.g. it is possible that a given covariate has a negligible effect on the 0.5-th quantile and a large effect on the 0.9-th quantile
- The methodology of quantile regression is well established for continuous outcomes, whereas *the extension to count data is not trivial*
  - main difficulty: the conditional quantile function of a discrete random variable cannot be a continuous function of the regression parameters
- We rely on Machado and Santos Silva 2005 → smoothing the counts through jittering in order to obtain a continuous working variable

jittered continuous variable Z =

count variable Y + uniform [0,1) random variable U

- Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.
- Machado, J. A. F. and Santos Silva, J. M. C. (2005). Quantiles for counts, *JASA* 100, 1226-1237.

Quantile regression for counts /cont.

jittered Z = count Y + uniform U

count	uniform	jittered
3	0.12	3.12
7	0.81	7.81

- Need a monotone transformation of the conditional quantile function of Z → similarly to a GLM for counts we choose a log function

$$Q_Z(\tau \mid \mathbf{x}, Y > 0) = \tau + \exp(\mathbf{x}' \boldsymbol{\beta}(\tau))$$

τ is the quantile order, e.g. τ=0.5 for the median

- The conditional quantile function of the count variable Y is

$$Q_Y(\tau \mid \mathbf{x}, Y > 0) = \lceil Q_Z(\tau \mid \mathbf{x}, Y > 0) - 1 \rceil$$

$\lceil t \rceil$ is the ceiling function (smallest integer $\geq t$)

- Estimation: separately for each quantile order (e.g. 0.10, 0.25, …), the regression coefficients are estimated through the linear quantile algorithm applied to the transformed jittered variable Z (estimators are consistent and asymptotically normal)

Quantile regression for counts /cont.

- Software for estimation: we used the Stata command **qcount** by A. Miranda (there is also an R package)
- Following Machado and Santos Silva, to average out the random noise we repeat jittering 1000 times and compute the average-jittered estimator (proved to be more efficient) – easily done with the software
- We report the results as ***partial effects*** of the covariates on the jittered variable *Z*, namely we consider the change in the following quantity:

$$Q_Z(\tau \mid \mathbf{x}_*, Y > 0)$$

$\mathbf{X}_*$ = covariates fixed
at the mean (if continuous)
or at zero (if binary)

- Partial effect: derivative for a continuous covariate, discrete change for a binary covariate
- Remark: $Y \in \{1,2,\ldots,20\}$ thus to recover the original scale of the credits we report the partial effects on *Z* multiplied by 3

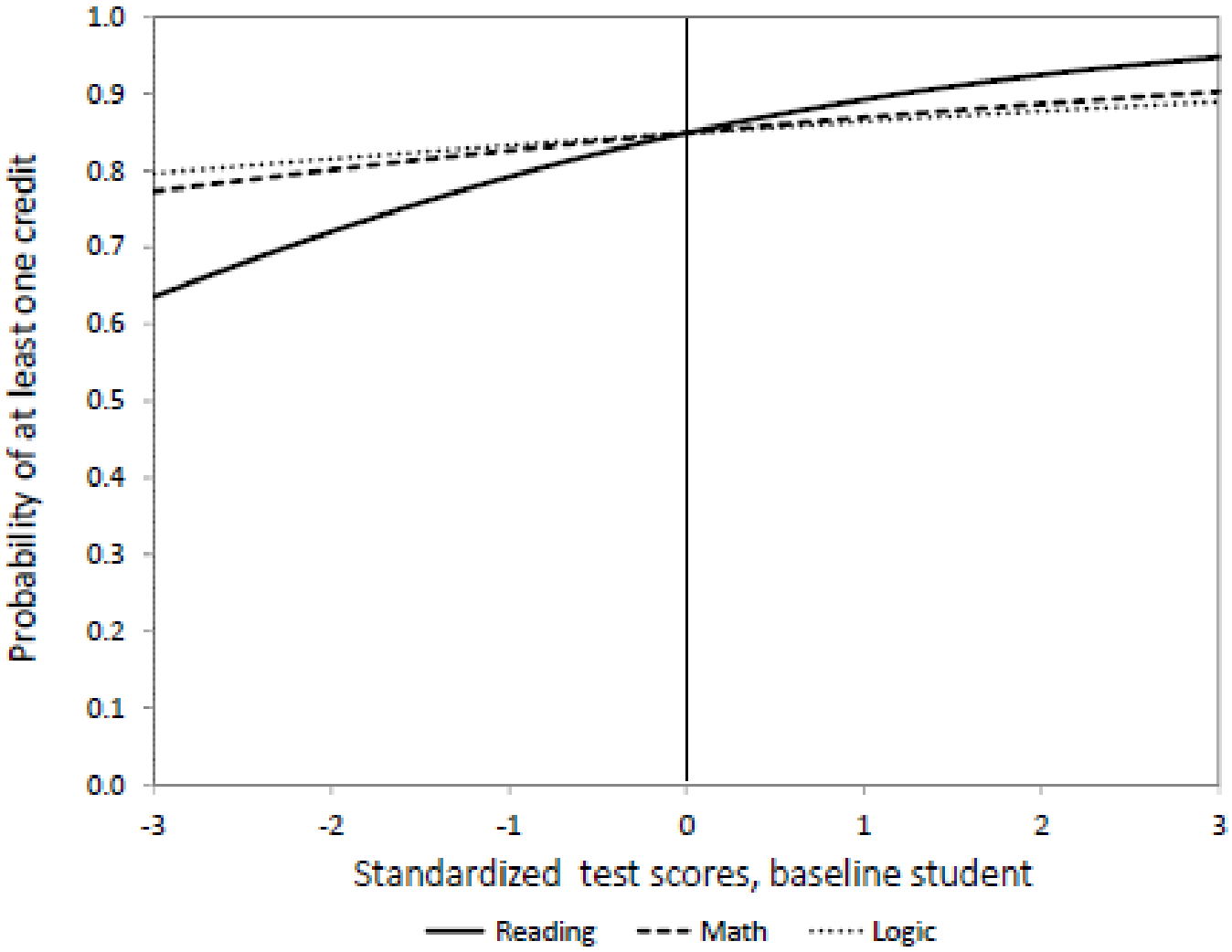
	Logit model	Quantile regression for τ∈{0.10,0.25,0.50,0.75,0.90}				
	$P(credits > 0)$	$f(credits \mid credits > 0)$				
		$Q(0.10)$	$Q(0.25)$	$Q(0.50)$	$Q(0.75)$	$Q(0.90)$
Prediction						
Baseline student†	0.849	12.963	19.854	31.874	44.252	48.971
Partial effect						
Female	0.016	*3.129*	*3.377*	2.182	0.035	0.939
	(0.026)	(1.241)	(1.568)	(2.278)	(1.514)	(1.121)
Far-away resident	-0.026	0.094	-1.482	0.888	0.210	-0.053
	(0.054)	(1.688)	(2.377)	(3.632)	(3.086)	(2.025)
HS type (ref: science)						
Humanities	-0.019	-3.512	-0.939	-1.135	-0.618	-0.435
	(0.062)	(2.182)	(9.170)	(2.741)	(2.958)	(2.622)
Technical	-0.048	-0.060	1.030	-2.475	-3.619	-2.118
	(0.036)	(1.152)	(1.671)	(2.483)	(2.609)	(2.186)
Other	-0.047	-2.386	*-3.749*	-5.682	*-8.179*	*-6.749*
	(0.048)	(1.663)	(1.715)	(4.179)	(3.797)	(2.587)
HS irregular career	*-0.177*	*-2.550*	*-6.936*	*-10.222*	-11.710	-3.589
	(0.056)	(0.924)	(1.609)	(2.083)	(7.658)	(4.852)
HS grade (+10)	*0.035*	*4.739*	*11.246*	*10.667*	*9.420*	*3.020*
	(0.008)	(0.934)	(0.918)	(1.258)	(1.469)	(0.785)
Std test scores						
Logic	0.016	0.278	1.397	0.443	-0.557	-0.279
	(0.014)	(0.718)	(0.771)	(0.939)	(0.733)	(0.637)
Reading	*0.050*	-0.018	-0.238	0.155	0.378	0.555
	(0.016)	(0.649)	(0.754)	(1.034)	(1.070)	(0.806)
Mathematics	0.021	0.686	*1.928*	*3.045*	*2.909*	*2.361*
	(0.018)	(0.786)	(0.893)	(1.190)	(1.091)	(1.058)

Baseline student: male, resident in Florence or surrounding provinces, HS scientific, HS regular career, mid-point HS grade (80), test scores at mean values (0).

From the logit model:

probability to get at least one credit by test scores

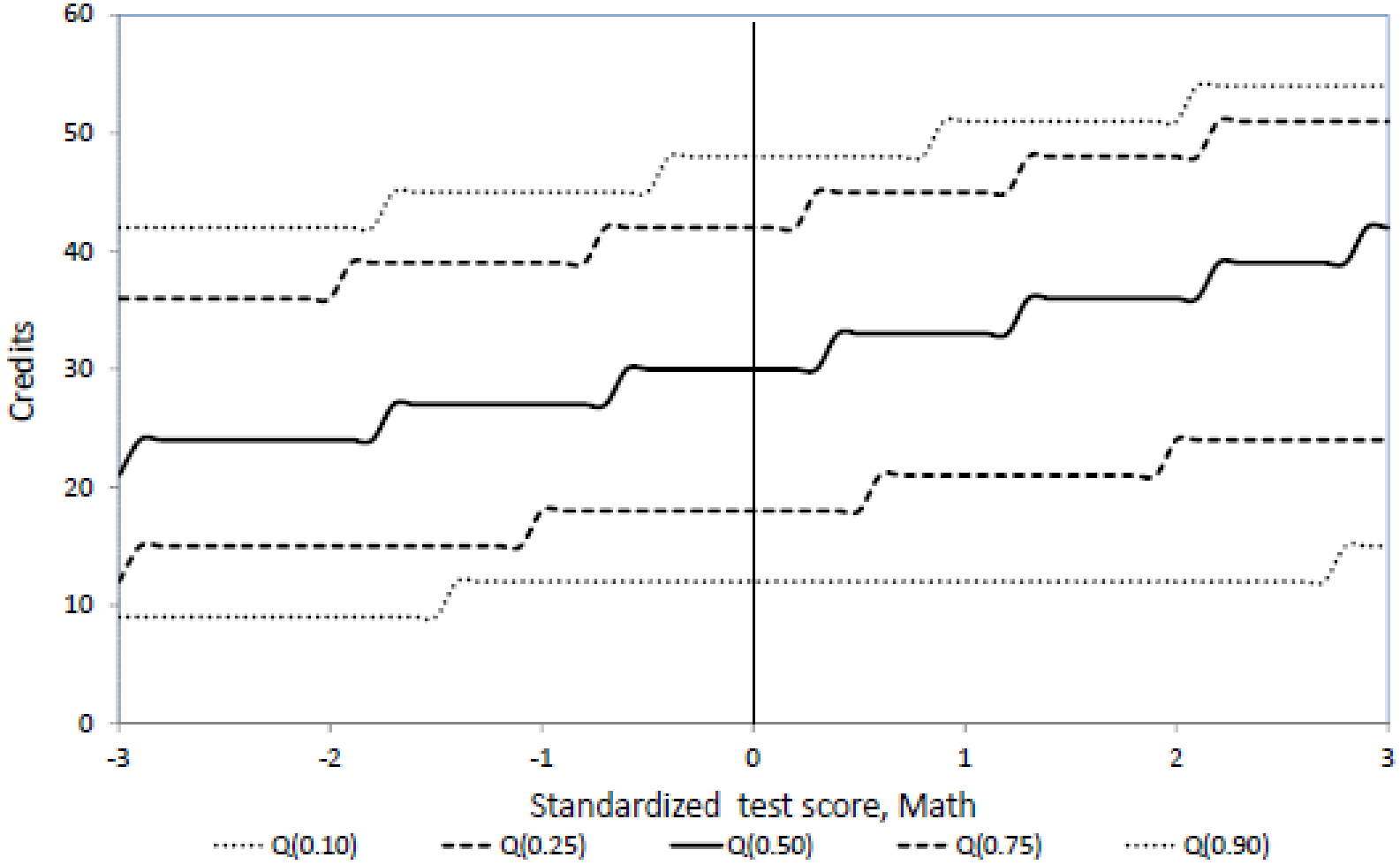
(baseline student)



From quantile regression on positive credits:

estimated quantiles as a function of the Math test score

(baseline student)



Assessing model fit

- In linear quantile regression, local model fit for each quantile τ can be evaluated through the $R^1(\tau)$ measure defined by Koenker and Machado (1999):

$$R^1(\tau) = 1 - \frac{V_{\text{model=full}}(\tau)}{V_{\text{model=null}}(\tau)}$$

where

$$V_{\text{model}}(\tau) = \sum_{i:y_i \geq \hat{y}_i(\tau)} \tau \left| y_i - \hat{y}_i(\tau) \right| + \sum_{i:y_i < \hat{y}_i(\tau)} (1-\tau) \left| y_i - \hat{y}_i(\tau) \right|$$

$$\hat{y}_i(\tau) = \left\lceil \tau + \exp(\mathbf{x}' \, \hat{\boldsymbol{\beta}}(\tau)) - 1 \right\rceil$$

- In quantile regression for counts $V_{\text{model}}(\tau)$ is not the objective function, however $R^1(\tau)$ still has an R^2-like interpretation so we propose to use it to assess model fit

Assessing model fit /cont.

$$R^1(\tau) = 1 - \frac{V_{\text{model=full}}(\tau)}{V_{\text{model=null}}(\tau)}$$

- In the application the fit is similar for all the considered values of τ

τ	0.10	0.25	0.50	0.75	0.90
$R^1(\tau)$	0.124	0.167	0.161	0.149	0.170

- The values of $R^1(\tau)$ are in line with those usually found in applications of linear quantile regression

Predictions

- The model can be used to predict the number of gained credits for a hypothetical student
- Many ways of making predictions, we chose the following
  - point prediction: median
  - interval prediction: interquartile interval
- Problem: we fitted quantile regression for the conditional distribution $Y|Y>0$ but we wish to predict a quantile of the marginal distribution Y → we developed a procedure to account for the hurdle part (see the paper)
- For example, for the baseline student
  - Median = 27
  - Interquartile interval = [12,42]

Substantive conclusions

- The results of the quantile regression confirm the findings of the concomitant-variable mixture model about predicting gained credits:
  - usefulness of background covariates
  - additional information (limited) yielded by the pre-enrolment test
  - higher score on Reading → higher probability of gaining at least one credit
  - higher score on Math → higher number of credits during the first year

Remarks on the method

- The hurdle quantile regression for counts has several merits:
  - *hurdle* → it models the probability of zero credits separately from the positive part of the distribution of credits
  - *quantile regression* → it avoids distributional assumptions + it allows to analyze the effects of covariates at different quantiles
  - *for counts* → it accommodates the discrete nature of gained credits
- Extending quantile regression to count data is not trivial:
  - we used the jittering approach of Machado and Santos Silva (2005). The noise induces a perturbation, which is *proportionally larger for small counts*. However, in our application the effect on the estimated quantiles is likely to be negligible, since the estimator is averaged over 1000 replicates, and the lowest considered quantile is the 10^{th} of the distribution of positive counts
  - worth to investigate alternative approaches avoiding jittering, such as Chanialidis, C., Evers, L., Neocleous, T. (2014). Bayesian density regression for count data. arXiv:1406.1882

Remarks on the method /cont.

- The modelling strategy of *hurdle quantile regression for counts* proved to be simple and effective → valuable also for other applications with zero-inflated count data
- Even in case of zero-inflated data, it is possible to omit the hurdle structure and apply quantile regression to the whole sample including the zeroes (see the simulation example of Machado and Santos Silva, 2005)
- However, we have a special interest in the zeroes (students who do not gain any credit), thus we prefer the hurdle specification that allows us to ***explicitly model the probability of having a zero*** (gaining zero credits)
  - On the contrary, a quantile regression model on the whole sample would yield results at a fixed grid (e.g. 0.10, 0.25, . . .) which do not allow to directly compute P(Y = 0 | x).

Mixtures vs quantile regression

- For the analysis of gained credits we used two approaches:
  - Concomitant-variable binomial mixture model
  - Hurdle model with quantile regression for counts
- The substantive conclusions from the two approaches are similar
- Both models are complex → need to convert model parameters into easily interpretable quantities (plots are helpful), even if we think *quantile regression* is more intuitive
- Both methods have controversial issues:
  - *mixture modelling* has to face the problem of choosing the number of components, in addition to several well-known issues related to the likelihood (multimodality, non-identifiability … see Larry Wasserman's blog 'Normal Deviate' where he concludes "*I have decided that mixtures, like tequila, are inherently evil and should be avoided at all costs*")
  - *quantile regression* requires to select a set of quantiles, there are well-known problems such as quantile crossing, our count data implementation is based on adding random noise (jittering) …

Mixtures vs quantile regression /cont.

- *Mixture modelling* has several motivations but only number 1. is relevant in our application on gained credits (where the outcome is unidimensional) :
  1. increasing flexibility
  2. summarizing a complex multivariate structure
  3. classifying units
- Given the aim of increasing the flexibility, *quantile regression* is preferable as it is simpler for model selection and fitting, and for interpreting the results

Thanks for your attention!

grilli@disia.unifi.it, rampichini@disia.unifi.it