Quantile regression coefficient modeling for counts to evaluate the productivity of university students

Modellazione dei coefficienti di regressione quantile per dati di conteggio per valutare la produttività degli studenti universitari

Viviana Carcaiso and Leonardo Grilli

Abstract The extension of quantile regression to count data raises several issues. In this research we compare a solution which exploits jittering to obtain a continuous working variable to a more recent approach, in which the coefficients of quantile regression are modeled by parametric functions. Both methods are applied for evaluating the effect of remote teaching on university students' productivity. In this context, the latter approach is found to be advantageous.

Abstract *L'estensione della regressione quantile ai dati di conteggio provoca diversi problemi. In questa ricerca confrontiamo una soluzione che sfrutta il jittering per ottenere una variabile continua con cui lavorare con un approccio più recente, in cui i coefficienti di regressione quantile sono modellati da funzioni parametriche. Entrambi i metodi sono applicati per valutare l'effetto della didattica a distanza sulla produttività degli studenti universitari. In questo contesto, l'approccio più recente risulta vantaggioso.*

Key words: COVID-19, jittering, parametric modeling, quantiles, university credits

1 Introduction

Quantile regression has the main advantage of being completely distribution-free and avoiding some of the restrictive assumptions of conventional regression: it does not require homoscedasticity or a specific type of distribution for the response and it

Viviana Carcaiso
Department of Statistical Sciences, University of Padova, Padova, Italy. e-mail: viviana.carcaiso@phd.unipd.it

Leonardo Grilli
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy. e-mail: leonardo.grilli@unifi.it

is very flexible since regression equations at different quantiles are fitted separately. Most of its theoretical developments and empirical applications concern continuous outcomes. Indeed, the response variable is generally assumed to be sampled from an absolutely continuous population, which is not true if the outcome is a count. The extension of quantile regression to count data raises several issues, mainly related to the fact that the conditional quantile function of a discrete random variable cannot be a continuous function of the regression parameters [6]. The solution suggested by Machado and Santos Silva (2005) [6] is to generate an artificial continuous variable by adding a uniform random variable to the original counts, a procedure referred to as jittering. This allows to obtain a continuous working variable whose quantiles have a one-to-one relation with those of the response (see [4] for an application to the productivity of university students). A recent alternative approach considered in this research is the one by Frumento and Salvati (2021) [3], who suggest applying the quantile regression coefficients modeling (QRCM) paradigm described in [2] to model a discrete response. As Machado and Santos Silva, they want to impose some degree of smoothing to the assumed distribution, without altering the response itself. Their solution permits to avoid jittering and was shown to provide numerous advantages, including parsimony, efficiency and simplicity in terms of interpretation. Here, we apply both methods to evaluate the effect of remote teaching due to COVID-19 pandemic on university students' productivity of freshmen from the academic year 2019/2020 of the University of Florence. The rest of the paper is organised as follows. Section 2 outlines the theory of the two methods, Section 3 illustrates the case study and the applied statistical models, and in Section 4 we discuss the obtained results.

2 Methods

In 2005, Machado and Santos Silva [6] proposed analyzing count data using quantile regression, which permits avoiding strong parametric assumptions and enables investigating every aspect of the conditional distribution of the response variable, not just its mean. However, there are some complications, related to the non-smoothness of the objective function combined with the discreteness of the response variable. The solution of Machado and Santos Silva [6], that will be referred to as QRCJ (Quantile Regression for Counts based on Jittering), is based on the artificial smoothing of the data using jittering.

More specifically, after defining $Z = Y + U$, where U is a uniform random variable independent of Y and $\boldsymbol{x}$, and noting that $Q_Z(p|\boldsymbol{x})$ is bounded from below by p, the authors specified a model of the form $Q_Z(p|\boldsymbol{x}) = p + \exp(\boldsymbol{x}'\boldsymbol{\beta}(p))$. In other words, they propose to use a monotone transformation of the conditional quantile function of Z, namely $T(Z;p) = \log(Z-p)I(Z > p) + \log(\zeta)I(Z \leq p)$, where ζ is an arbitrary small positive number, and then $\boldsymbol{\beta}(p)$ can be estimated by running a linear quantile regression of $T(Z;p)$ on $\boldsymbol{x}$. The value of $\hat{\boldsymbol{\beta}}(p)$ depends not only on the sample information but also on the specific realization of U, thus the authors

propose to generate m "jittered" samples and to average the estimates (they call this procedure "average-jittering"). The average jittering estimator is proved to be more efficient than the one based on a single sample and, under mild conditions on the covariates, to be consistent and asymptotically normal, therefore standard asymptotic theory can be used to perform inference. The simulation study in [6] shows that the proposed estimator has good properties in finite samples of size $n = 500$.

The fact that quantiles are estimated one at a time and that no parametric structure is assigned to the coefficient functions has relevant drawbacks [3]. The quantile regression coefficients modeling (QRCM) approach, first developed by [2], and then applied to count data by [3], consists in describing the regression coefficients by smooth parametric functions with closed-form mathematical expressions and should provide a gain in terms of efficiency and simplicity of interpretation. In the general QRCM framework presented by [2] the q regression parameters are defined as $\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$, where $\boldsymbol{b}$ is a vector of k known functions of p that are assumed to be continuous and differentiable, and $\boldsymbol{\theta}$ is a $q \times k$ matrix. In the specific case of quantile regression for count data, the model proposed by Frumento and Salvati is driven by the empirical evidence that the estimators obtained by applying QRCM to the jittered response, $Z = Y + U$, and directly to $Y^{\circ} = Y + E[U]$ are almost identical, due to imposed parametric structure [3]. Therefore, after assuming, without loss of generality, $E[U] = 0.5$, the model is applied to a transformation $T(\cdot)$ of $Y^{\circ} = Y + 0.5$, that is

$$Q_{T(Y^{\circ})}(p|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\theta}\boldsymbol{b}(p).$$

The estimation is carried out by minimizing the integrated objective function, which is the integral, with respect to the order of the quantile, of the loss function of standard quantile regression. This estimation approach allows to estimate the entire quantile process instead of a discrete set of quantiles. Furthermore, the integrated loss function is a smooth function of its arguments, and this allows to carry out minimization using standard algorithms, like Newton–Raphson or gradient search, and to apply the standard theory of M-estimators to investigate the asymptotic behaviour of the estimator. Fitting linear quantiles is the most common practice because of the simplicity of the interpretation [2], but in [3] it is also proposed a logarithmic transformation.

The idea is that the imposed parametric structure should yield some gain in terms of efficiency, as shown by the simulation results reported by the authors in [3]. However, specifying a "good" parametric model for a quantile function it is not easy, since there are many different possible specifications. The model is determined by the choice of $\boldsymbol{b}(p)$, which must be defined in advance, and by the restrictions that are imposed on $\boldsymbol{\theta}$. In practice, the most common choices are polynomials, splines, piecewise linear functions, roots, logarithms, trigonometric functions, quantile functions of known distribution (e.g., that of a Normal, Beta or Gamma distribution), and combinations of the above. In many situations, some of the quantile regression coefficients can be assumed to be linear functions of p, or not to depend on p at all. [2] propose a goodness-of-fit test procedure that is based on the idea of assessing the

model fit by comparing the distribution of $F(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})$, $i = 1,\ldots,n$, with a $U(0, 1)$ distribution, but it just represents an approximation in the framework of count data.

3 Application

The evaluation of the impact of the change to remote teaching on university students' careers represents an interesting application of quantile regression for count data. The data were collected in the administrative records and concern first-year students who enrolled in the Bachelor's degree courses of Psychological Sciences and Industrial Design in the academic years 2018/2019 and 2019/2020 at the University of Florence. Two different degree courses are selected to have an insight in different fields of study. The idea is to compare the productivity of freshmen in the second semester of 2019/2020, which is the one affected by online teaching, to that of first-year students in the second semester of 2018/2019, who received the usual frontal lectures. Both cohorts attended regular lectures in the first semester, but were exposed to different forms of teaching in the second one. In both academic years of interest the study plan remained the same. After excluding students who got no credits in the first semester, the data set consist of 946 first-year students, whose characteristics are displayed in Table 1. We can notice that within each degree course the two cohorts have similar characteristics.

Table 1 Summary of background characteristics of freshmen by degree course and year of enrollment (2018 or 2019), University of Florence.

	Psychological Sciences			Industrial Design		
	2018	2019	Total	2018	2019	Total
Nr. observations	313	336	649	139	158	297
Gender (%)						
Female	76.7	82.1	79.5	72.7	62.0	67.0
Male	23.3	17.2	20.5	27.3	34.0	33.0
Type of HS (%)						
Scientific	37.7	32.7	35.1	32.4	25.3	28.6
Humanities	16.3	17.0	16.6	5.04	6.96	6.06
Language	7.35	7.14	7.24	2.88	3.80	3.37
Human Sciences	24.6	19.9	22.2	12.2	9.49	10.8
Art School	1.28	2.98	2.16	22.3	29.1	25.9
Technical	9.27	15.2	12.3	19.4	18.4	18.9
Other	3.51	5.06	4.31	5.76	6.97	6.40
HS grade						
Average	82.1	80.3	81.17	76.9	78.3	77.63
Std. error	10.9	11.1	11.00	10.1	11.1	10.79

Students' productivity is measured by the number of gained credits (ECTS). According to the study plan, students should obtain approximately 30 credits per semester. Since the number of credits is always a multiple of 3, the response variable used in the models, denoted by Y, is defined as the number of credits gained in the second semester divided by 3. Since Y presents an irregular distribution in

both degree courses, quantile regression represents an appealing methodology. The covariates included in the models are:

- X_1: number of credits obtained during the first semester, centred around 15.
- X_2: dummy variable for cohort 2019 vs 2018.
- X_3: dummy variable for male vs female.
- X_4: high school grade, centred around 80.
- $X_5, ..., X_{10}$: dummy variables for the types of high school (except for the baseline category "Scientific").

The analysis is performed separately for each bachelor's degree course and the effect of interest is represented by the estimate of the coefficient of X_2, since this is the variable that distinguishes the students whose second semester was affected by remote teaching (cohort 2019) from the others (cohort 2018). Both the QRCJ and QRCM approaches are exploited and, more specifically, the applied models are

$$\text{QRCJ}: \quad Q_Z(p|\boldsymbol{x}) = p + \boldsymbol{x}'\boldsymbol{\beta}(p) = p + \beta_0(p) + \sum_{i=1}^{10} \beta_i(p) x_i,$$

$$\text{QRCM}: \quad Q_{Y^\circ}(p|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \beta_0(p|\boldsymbol{\theta}) + \sum_{i=1}^{10} \beta_i(p|\boldsymbol{\theta}) x_i.$$

In both of them the transformation of the working variable is the identity function: in the QRCJ approach the logarithmic function is not applied, in order to ease the comparison with QRCM. The methods were computationally implemented by using the function `rq` of the R package `quantreg` [5] for QRCJ, and the function `iqr` of the `qrcm` package [1] for QRCM.

For both degree programs, a variety of parametric models were applied to the conditional quantile function of $Y^\circ = Y + 0.5$ and they were compared based on the number of free model parameters, the integrated loss function, the p-values of the Kolmogorov–Smirnov (KS) test for goodness-of-fit and the similarity of the estimates of the constant and $\beta_2(p|\boldsymbol{\theta})$ to the ones obtained from QRCJ. In both cases, in the selected model the intercept and the coefficient which represents the effect of interest, $\beta_2(p|\boldsymbol{\theta})$, are modelled in a more complicated and flexible way. In particular, they are formulated as combinations of shifted Legendre polynomials and logarithmic functions. The other coefficients are instead approximated by constant (for the ones related to the type of high school) and linear functions.

4 Discussion and final remarks

Quantile regression represents an appealing methodology for dealing with an outcome variable characterized by an irregular distribution, as in the case of the number of credits gained by university students. In this research two different methods for applying quantile regression to count data were employed: the jittering approach

of [6] and the parametric modelling one of [3]. The obtained results show that they lead to similar estimates of regression coefficients, especially for the variables which were modelled through flexible parametric functions (combinations of polynomial and logarithmic functions) in the QRCM setting. The QRCM approach in most cases entails a gain in efficiency, especially in the tails of the distribution of the response. This gain is generally more substantial for the control variables that are approximated by constant functions, whereas is lower or not present for the coefficients described by more flexible parametrizations. The parametric modeling approach has the advantage of providing estimates of the regression coefficients that are smooth functions of p, and makes the interpretation more straightforward. Moreover, it allows to estimate multiple quantiles at the same time, resulting in a much faster computation. However, model selection is not easy and requires time and expertise. The suggested goodness-of-fit procedure is not completely appropriate in the case of count data, and the integrated loss function is not very useful since it depends on the number of parameters. In our case study, the number of possible models was reduced by the fact that there is a specific coefficient of interest, but it was still quite large. A combination of QRCJ and QRCM was found to be a good method for model selection. Indeed, the estimates obtained with the jittering method can be used as a benchmark when choosing among a set of parametric specifications, in order to derive a satisfactory functional form for some specific covariates, keeping always under control the value of the loss function and the goodness of fit.

As far as the interpretation is concerned, the effect of remote teaching on Psychological Sciences freshmen is negative, but small, at most quantiles. It is stronger on the tails, namely for students with low or high productivity (but not for the ones with the greatest performance), however all the estimates are not significantly different from 0 (at a level of 5%). For freshmen in Industrial Design the effect is instead positive, but again small, and stronger for the group of the most productive students. Indeed, the QRCM estimates do not reach statistical significance at 5% except at a few quantiles around 0.95. In both courses the data size does not seem large enough for identifying such small effects.

References

1. Frumento, P. (2021). `qrcm`: quantile regression coefficients modeling. R package version 3.0. http://CRAN.R-project.org/package=qrcm.
2. Frumento, P. and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics*, 72: 74-84.
3. Frumento, P. and Salvati, N. (2021). Parametric modeling of quantile regression coefficient functions with count data. *Stat Methods Appt*.
4. 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. *Stat. Modelling* 16(1):47–66.
5. Koenker, R. (2021). `quantreg`: Quantile Regression. R package version 5.85. https://CRAN.R-project.org/package=quantreg.
6. Machado, J. A. F. and Santos Silva, J. M. C. (2005). Quantiles for counts, *J. Am. Stat. Assoc.*, 100, 1226–1237.