International Journal of Statistics and Systems
ISSN 0973-2675 Volume 13, Number 1 (2018), pp. 9-14
© Research India Publications
http://www.ripublication.com

Predicting Student Academic Achievement: A Quantile Regression Approach

Ilaria Lucrezia Amerise
Department of Economics, Statistics and Finance,
University of Calabria,Italy.

Abstract

The purpose of this paper is to apply the weighted quantile regression method to college admissions in order to consider the controversial behavior of the final mark of high school diploma (HSD). The new approach takes into account the fact that student characteristics have a varying impact at different part of the indicator of college performance (here: weighted average mark or WAM), that is, for particular quantiles. The main advantage of this method is the ability to handle models in which data exhibit dissimilar levels of quality. Furthermore, quantile regression is also a robust technique to outliers in the response variable. Our results reveal that there exists a complementary correlation between HSD and WAM and between WAM and the average grade of the first term of the last year at high school (AGFT).

Keywords: Quantile regression, Robust Estimation, Academic Achievement.

1. INTRODUCTION

A surge in university enrolment in the last decade has widened the gap between the number of students who enter the university system and the number who graduate. See, for example, Hanushek, E. (1996) and Mo-Yin, S.Tam *et al.* (2002). The aim of this paper is to show that, although the relationship between applicant characteristics and academic effectiveness may in fact be negative, nil, or positive according to ordinary least squares (OLS) estimation, such relationship may change sign or become significant along the conditional WAM distribution. To estimate it separately we use the quantile regression (QR). This method is a breakthrough in the econometric analysis. In fact, by estimating various conditional quantile functions, researchers are able to characterize the conditional distribution of interest and establish a complete description of there rationship between the response variable and the regressors. See Koenker & Bassett (1978), Koenker, (2005) and Hallock &

Koenker, (2001) for a good review. Quantile regression is a generalization of the least absolute deviation method of estimation. Assuming $(y_i, \boldsymbol{x}_i)$, i=1,2,⋯,n is a sample of pairs where y_i is the response variable (here: the WAM) and $\boldsymbol{x}_i$ is the vector of explanatory variables (here: HSD, AGFT, gender), the quantile regression model can be written as follows:

$$y_i = x_i^t \beta(\alpha) + u_i \qquad with \quad \mathrm{Q}_\alpha(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^t\, \beta(\alpha) \tag{1}$$

where $0 < \alpha < 1$, $\beta(\alpha) = [\beta_0(\alpha),\beta_1(\alpha),\cdots,\beta_m(\alpha)]$ are the regression coefficients and $\{u\}_{i=1}^n$ is a sequence of independent random variables that accounts for the erratic component in y_i that cannot be explained by $\mathbf{x}_i$. The coefficient β_j (α) can be interpreted as the marginal change in the α-th conditional quantile of the response variable with respect to α-th regressor x_j. The term ”marginal” indicates a tendency and does not imply that the observation $\mathbf{x}_i$ would still be in the α-th quantile if the corresponding value $x_{i,j}$ changed. Moreover, quantile regression estimates are reliable in presence observations that are far in the direction of the response variable and, as long as y_i remains on one of the side of the regression plan,the estimate β will remain unchanged. Quantile regression estimates are not guaranteed to be unique for the given level α (consider,for example, the interval of medians that occurs with an even number of univariate observations). A very interesting feature of the system of quantile regression functions (1) is that all the observations are employed in the estimations, but a different score: α or $(1-\alpha)$, is assigned to them depending on whether $y_i > \boldsymbol{x}_i^t\beta(\alpha)$ or $y_i \le \boldsymbol{x}_i^t\beta(\alpha)$. In this sense, the α-th regression quantile estimator for β solves the following problem:

$$\min_{\beta(\alpha)}\left\{\alpha \sum_{y_i > x_i^t\beta(\alpha)} [y_i - \boldsymbol{x}_i^t\beta(\alpha)] - (1-\alpha) \sum_{y_i \le x_i^t\beta(\alpha)} [y_i - \boldsymbol{x}_i^t\beta(\alpha)]\right\} \tag{2}$$

As α goes from 0 to 1, the entire distribution of the response y_i, conditional on $\mathbf{x}_i$ is described. For each level of α it can be shown whether the effect of a given regressor is positive or negative, and how large this effect is compared to other quantiles. Consequently, quantile regression is capable of providing insight into the center as well as the lower and upper tails of y | **x**. This analysis is particularly useful when the regressors have a different impact on different parts of the conditional distribution of the response.

2. WEIGHTED QUANTILE REGRESSIONS

Unfortunately, the robustness of quantile regression cannot be extended to the effect of observations that are distant in the space of regressors, where even a single point dragged far enough toward infinity, can cause that all quantile regression hyperplanes go through it. One unrealistic assumption underlying (2) is that each point in one side of the regression plan provides equally precise information about the deterministic part of the response variable. In situations when it may not be reasonable to assume that every observation should be treated equally, a weighting system can often be used to maximize the efficiency of parameter estimation. In the admissions literature it has been recognized that student characteristics not only vary along the entire distribution of a performance indicator, but also vary in terms of data quality. In this section, we attempt to deal with this problem by adjusting fit to give less weight to observations thought to be less reliable. Different systems of weights are discussed in Amerise, (2016). Weighted quantile regression works by incorporating extra non negative constants $v_i \geq 0$, i=1,2,⋯,n associated with each data point, into the fitting criterion.

$$\min_{\beta(\alpha)} \left\{ \alpha \sum_{y_i^* > \boldsymbol{x}_i^{*t}\beta(\alpha)} [y_i^* - \boldsymbol{x}_i^{*t}\beta(\alpha)] - (1-\alpha) \sum_{y_i^* \leq \boldsymbol{x}_i^{*t}\beta(\alpha)} [y_i^* - \boldsymbol{x}_i^{*t}\beta(\alpha)] \right\} \quad (3)$$

with y_i^* =$w_i^{-0.5}$ y_i ,$\mathbf{x}_i^*$ = $w_i^{-0.5}$ $\mathbf{x}_i$, $\boldsymbol{w}$ = $\mathbf{v}$ ($-t_\alpha$), where ($-i$) means that the i-th case is being omitted from the set and t_α denotes the set of observations with zero weight. The size of the weight w_i quantifies the reliability of the information contained in the i-th observation. It is important to note that the weight for each observation is given relative to the weights of the other observations; so different sets of weights can have identical effects. In addition, since w_i must be positive, it may happen that different weighting schemes $v_i \geq 0$, i = 1,2,⋯,n involve different subsets of the data set.

In order to solve the weighted problem 3, it is sufficient to divide the response and the vector of explanatory variables by the square root of its weight and to solve the unweighted problem with the couple (y_i^* , $\mathbf{x}_i^*$), i=1,2,⋯,n thus obtained. The theory behind this method is based on the assumption that the weights are fixed and known in advance. This assumption rarely holds so estimated weights must be used instead. In the OLS framework an iterative estimation procedure is invoked and a similar approach could be considered for the QR estimation. Iterative methods typically converge after only a few iterations. To keep the over all computation effort at a tractable level, we used only a single iteration (one-step estimation) in the hope that the resulting estimators would have good statistical properties. Suppose that we have an estimator $\widetilde{\beta}$ of β in (3) ; let $r_i = y_i - \boldsymbol{x}_i^t\widetilde{\beta}$ denote the residuals and let s =1.48261+5/(n−2) [*median* (r_i^2)] be an empirical measure of the scale. Our strategy proceeds by applying the unweighted QR to the original data, down weighting observations with discordant residuals, deleting observations with extreme residuals, proceeding with a standard QR on the new data set. Choosing a weighting method is a non trivial problem, with no single method dominating all others. In this sense, we have tried fifteen weight functions (Huber, Andrews, Hampel, Tukey, Welsch among the others) obtaining the best results with the following scheme v_i = 2[1+ exp (|r_i|/2.985s)]$^{-1}$. To enable the comparison of different quantile regressions, robust measures of performance are necessary. One such measure is R_1, the correlation between the original data $\mathbf{y}$ and the predicted vector $\widetilde{\boldsymbol{y}}$. It is a typical measure of linear concordance between observed and predicted data often used to quantify the fitting in non standard regression problems. The model fits well the observed data for values of *cor* $(\boldsymbol{y}, \tilde{\boldsymbol{y}})$ near 1, while the fit is poor when *cor* $(\boldsymbol{y}, \tilde{\boldsymbol{y}})$ tends to 0.

3. RESULTS

Given that we were not able to find any accurate account of the dropouts, our study objective was to focus on students still enrolled at the moment of the survey. See, for example ,(Hoxby, 2000). Restriction to those students who attend regularly the courses determines a serious selection bias, partially weakened by the assumption that students with a low academic preparation can be considered a good proxy for dropouts. A sample of 105 Students (50 females and 55 males) was collected over a span of one month at campus Arcavacata (University of Calabria) during the 4th trimester, 2017. The questionnaire included several questions, but only three items were deemed useful to explain the WAM. In particular, we regressed WAM against HSD (x_1), AGFT (x_2) and gender (x_3). The correlation between HSD and AGFT is 0.21 indicating that both could then be used in an additive fashion to provide separate pieces of information, both of which would contribute to explain the WAM. The quantile regression coefficients for deciles 0.1 to 0.9 are presented in Table (1). For each α, $\boldsymbol{x}^t_i \beta(\alpha)$ defines the α-th sample quantile of $y|\mathbf{x}$. This means that, if the whole population of $y|\mathbf{x}$ were sampled, then α% of the values would be below of $\boldsymbol{x}^t_i \beta(\alpha)$ and $(1-\alpha)$% above. The first four columns of Table (1) report the coefficients whereas column 5 to 8 report the corresponding t-Student's calculate with the procedure *nid* of the R package quantreg. The last column in the table shows the model adequacy. Also included is the least squares estimate. Our findings reveal that HSD (β_1) has a positive effect on WAM and it is moderately significant across the nine quantiles (the effect is negative and not significant for OLS). The impact is particularly high for the lowest quantiles (0.1 and 0.2) suggesting that HSD is linked to poor-grade students rather than academic successful experiences.

Table 1: Quantile regression coefficients

α	β_0	β_1	β_2	β_3	$t(\beta_0)$	$t(\beta_1)$	$t(\beta_2)$	$t(\beta_3)$	R_1
0.1	2.857	0.075	1.858	0.358	1.416	4.033	7.273	0.688	0.128
0.2	-4.354	0.087	2.585	1.586	-5.873	4.231	13.747	2.555	0.202
0.3	4.334	0.028	2.356	1.480	1.775	1.261	7.620	3.068	0.116
0.4	3.837	0.020	2.489	2.849	13.804	1.487	14.620	20.378	0.316
0.5	3.809	0.077	2.005	0.856	1.138	2.319	4.803	1.160	0.355
0.6	3.486	0.053	2.381	0.932	5.674	2.380	7.190	1.284	0.487
0.7	5.549	0.063	2.212	0.920	6.214	1.797	4.925	1.147	0.754
0.8	9.649	0.069	1.521	1.593	3.147	3.301	4.488	2.796	0.955
0.9	13.086	0.062	1.277	1.125	5.038	2.690	8.120	1.653	0.982
Ols	20.034	-0.009	0.898	-0.591	10.851	-0.639	4.465	-1.708	0.499

On the other hand, all the quantiles (included OLS) confirm the importance of AGFT (β_2) as a predictor of student outcomes at university across the quantiles of the conditional WAM distribution. The coefficient β_2 is always positive, as expected, and highly significative.

Gender has only very slight effect on WAM with the exception of the lower deciles. For male students the intercept is $\beta_0 + \beta_3$ indicating that males of our samples tend to obtain greater values of WAM. Apparently, a gender impact was not found for OLS.

The goodness of fit R_1 is very near to one for the two highest deciles meaning that our model is more appropriate for lover achiever students than under performer students. This finding may imply that other factors, perhaps the type of secondary school or the hours spent studying could have greater influence on the right side of the WAM distribution.

4. CONCLUSIONS AND FUTURE RESEARCH

This paper has investigated the role that the high school performance of a student plays in undergraduate admissions to university, by using a system of quantile regressions. The following are the main research contributions of this paper. The final mark of high school diploma and the average grade of the first term of the last year at high school are important *a priori* determinants of university outcomes. However, their impact appears to be more significant for the highest deciles of our performance indicator (weighted average mark for university courses) than for low-achieving students. A selection index based on these two variables could result technically inaccurate because did not show clear discriminant relationships to score and compare all the applicants. The classical method of OLS has not been able to detect some important aspects of this relationship. Two potential directions for future research should be considered: to devise an iterative mechanism for weighted quantile regression for obtaining a greater accuracy on estimates and to establish a specific measure of goodness of fit for various quantile regressions on the same data.

REFERENCES

[1] Amerise, I.L. (2016). “Iteratively reweighted constrained quantile regressions”. Advances and Applications in Statistics, Vol 49, Issue 6, 417 – 441.

[2] Hanushek, E. (1996). “School resources and students performance”. In: Burtless, G. (Ed.) Does Money Matter? The effect of School Resources on Student Achivement and Adult Success. Brookingd Institution, Washington, D.C., 43-73.

[3] Hallock, K. and R. Koenker: “Quantile Regression” (2001). Journal of Economic Perspectives, Vol 15 (4), 143-156.

[4] Hoxby, C. (2000). “The Effects of Class Size on Student Achievement: New

Evidence from Population Variation”. The Quarterly Journal of Economics, Vol 115 (4), 1239-1285.

[5] Koenker, R. and Bassett G., (1978). “Regression quantiles”. Econometrica. Vol 46, 33-50.

[6] Koenker, R. (2005). “Quantile Regression”. Econometric Society Monograph Series, Cambridge University Press.

[7] Koenker, R. (2013). “Quantreg: quantile regression” R package. Available at http://CRAN.R-project.org/package=quantreg

[8] Mo-Yin S.Tam, Gilbert W. Bassett Jr. and Uday Sukhatme (2002). “New Selection Indices for University Admission: A Quantile Approach”. Statistics in indistry and Tecnology: Statistical data analysis, 67-76.