ISSN 1648-8776
JAUNŲJŲ MOKSLININKŲ DARBAI. NR. 1 (41). 2014

APPLICATION OF QUANTILE REGRESSION FOR THE PREDICTION OF STUDY RESULTS: THE CASE OF KTU

Laura Viselgaitė, Audrius Kabašinskas
Kaunas University of technology, Kauno Kolegija

Introduction

The idea of this paper is to forecast the final grade of module Mathematics 1 studied at the Kaunas University of Technology (hereinafter referred to as "KTU") in the beginning of the semester. Three criteria are selected as forecast factors: students' gender, national mathematics exam result (hereinafter referred to as "VBE") and the result of the university mathematics test (hereinafter referred to as "test"), which is taken by each student in the beginning of the semester. There were 217 students selected from the KTU faculty of Economics and Management, who have studied Mathematics 1 through 2009/10 and 2010/11. During the test, at the beginning of semester, each student must solve 15 simple problems from the school mathematics. These problems include operations with fractional expressions, operations with exponential and power type expressions, operations with root and irrational expressions, solving inequalities, linear and quadratic equations, description of given functions and their graphs, the domain of a function, and derivatives.

Mathematics 1 includes such topics as matrices, linear optimization, vectors, analytic geometry in space R^2 (line, second order curves), basics of mathematical analysis (sets, series, limits), calculus of one and two variables and basics of integration. Criteria directly related to the final grade, i.e. three semester assignments and an exam, are not included in our research.

Multiple linear regression models show the relationship between the dependent variable (in this case – final grade) and explanatory variables by mean. Using a quantile regression approach, it is allowed to choose for which quantile to make the model. In this research we have made a linear regression model and five quantile regression models. All models were analyzed and compared to determine the best result.

The aim of the article – to find out how accurately we can predict the final grades of Mathematics 1 module for each student in the beginning of the semester.

Research objective – to find out criteria which are not directly related to the final grade (to make a final grade prediction model without module Mathematics 1 semester and exam results).

Research methods – quantile and linear regression models.

Research results – estimated model for final Mathematics 1 results forecast at KTU.

Data for the research

In this research we use four variables: final grade, students' gender, VBE and test. Students' gender is a pseudo variable where code 1 means female and 0 means male gender. In our data we can see that more female students have chosen the KTU faculty of Economics and Management. We have 64.98 percent female students and 35.02 percent male students. VBE results are evaluated in percent – from 0 to 100. The mean of this variable is 64.73 percent and the standard deviation is 26.8. The test results are estimated in a ten-point system – from 0 to 10. The average test result is 6 where the minimum is 0.5 and maximum 10, while the standard deviation is 2.25. The final grade is also evaluated in a ten-point system. Research variables are not normally (Gaussian) distributed, because a minor part of students get low marks or bad exam results. All criteria correlations are positive and significant. The final grade has the strongest correlation with "test" variable – 0.74.

Outliers are also very important in every research. Data was checked by the following criteria: standardized residual, studentized residual, centered leverage, CooksD and Std. DfFit. If two from five criteria identified observed value as an outlier, it was removed from research. After the outliers test, 15 students' results were removed (202 remaining) from the data set.

Methodology

Quantile regression model

A more comprehensive picture of the effect of the predictors on the response variable can be obtained by using Quantile regression. Quantile regression models are the relation between a set of predictor variables and specific percentiles (or quantiles) of the response variable. It specifies changes in the quantiles of the response. For example, a median regression (median is the 50^{th} percentile) of infant birth weight on mothers' characteristics specifies the change in the median birth weight as a function of the predictors. The effect of prenatal care on median infant birth weight can be compared to its effect on other quantiles of infant birth weight [4]. In a linear regression, the regression coefficient represents the increase in the response variable produced by a one unit increase in the predictor variable associated with that coefficient. The quantile regression parameter estimates the change in a specified quantile of the response variable produced by a one unit change in the predictor variable. This allows for comparing how some percentiles of the Lahontan cutthroat trout weight may be more affected by certain stream characteristics [2] than other percentiles. This is reflected in the change of the size of the regression coefficient. Quantile regression applications help to explore food expenditures dependence from household incomes in more detail [5] or evaluate the expected performance of a risky asset and forecast its expected return [7].

Let *Y* be real valued random variable, characterized by distribution function

$$F(y) = \Pr(Y \leq y) \quad (1)$$

Then for any $\tau \in (0,1)$, the τ *th* quantile of *Y* is defined:

$$Q(\tau) = inf\{y: F(y) \geq \tau\} \quad (2)$$

For example, the median is $Q\left(\frac{1}{2}\right)$, and the first quartile $Q\left(\frac{1}{4}\right)$. Quantile function (2) fully characterizes variable *Y* (in the same sense as distribution function *F*) [6].

For a random sample $\{y_1, \ldots, y_n\}$ of *Y*, the general τ *th* sample quantile $\xi(\tau)$, which is the analog of $Q(\tau)$, is formulated as a minimizer:

$$\xi(\tau) = \arg min_{\xi \in R} \sum_{i=1}^{n} \rho_\tau(y_i - \xi) \quad (3)$$

where $\rho_\tau(z) = z(\tau - I(z < 0)), \tau \in (0,1)$ and where *I* denotes the indicator function. The loss function ρ_τ assigns a weight of τ to positive residuals $y_i - \xi$ and a weight of $1 - \tau$ to negative residuals. Using this loss function, the linear conditional quantile function extends the τ *th* sample quantile $\xi(\tau)$ to the regression settings in the same way that the linear conditional mean function extends the sample mean. Recall that Ordinary Least Squares (OLS) regression estimates the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving for

$$\hat{\beta} = \arg min_{\beta \in R} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \quad (4)$$

The estimated parameter $\hat{\beta}$ minimizes the sum of squared residuals in the same way that the sample mean $\hat{\mu}$ minimizes the sum of squares:

$$\hat{\mu} = \arg min_{\mu \in R} \sum_{i=1}^{n} (y_i - \mu)^2 \quad (5)$$

Likewise, quantile regression estimates the linear conditional quantile function $Q(\tau|X = x) = x'\beta(\tau)$, by solving

$$\hat{\beta} = \arg min_{\beta \in R} \sum_{i=1}^{n} (y_i - x_i'\beta) \quad (6)$$

for any quantile $\tau \in (0,1)$. The quantity $\hat{\beta}(\tau)$ is called the τ *th* regression quantile. The set of regression quantiles $\{\beta(\tau): \tau \in (0,1)\}$ is referred to as the quantile process [1]. SAS [3], [8] software procedure "quantreg" is developed for quantile regression. Quantile regression may be calculated for any given quantile. In most cases researchers use 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles.

Multiple linear regression model

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. Formally, the model for multiple linear regression, with given data set $\{y_i, x_{i1}, \ldots, x_{ip}\}$ for $i = \overline{1,n}$, is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad for\ i = \overline{1,n} \quad (7)$$

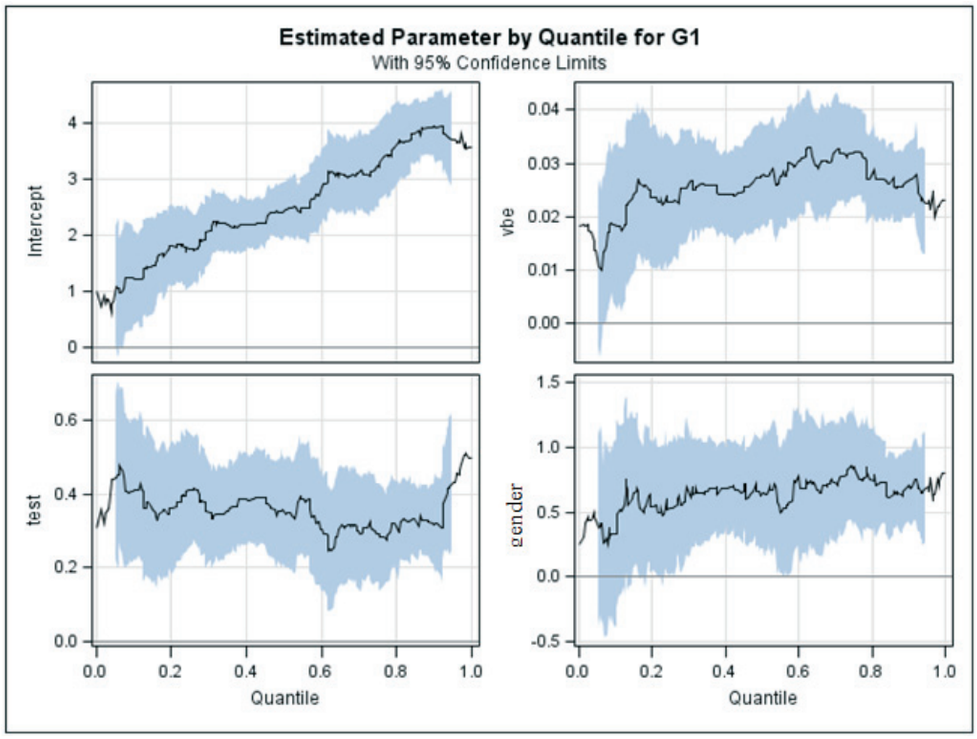

Fig. 1. Quantile regression process plots with 95 percent Confidence Limits

where β_0 – intercept, – parameters and ε_i – error variable [9]. The multiple linear regression model was selected using the stepwise method. This method specifies that variables are selected for the model based on a stepwise – regression algorithm, which combines forward – selection and backward – elimination steps. This method is a modification of the forward – selection method in that variables already in the model do not necessarily stay there.

Results

In this research we have used SAS [3], [8] software procedure "quantreg". Quantile regression was calculated for 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles. Let the variable "final grade" hereinafter be referred to as "G1" in formulas. The quantile regression equations are following:

$$G1(0{,}05) = 0{,}80 + 0{,}014 * vbe + 0{,}44 * test + 0{,}44 * gender$$
$$G1(0{,}25) = 1{,}75 + 0{,}023 * vbe + 0{,}409 * test + 0{,}524 * gender$$
$$G1(0{,}50) = 2{,}40 + 0{,}028 * vbe + 0{,}356 * test + 0{,}631 * gender$$
$$G1(0{,}75) = 3{,}22 + 0{,}032 * vbe + 0{,}293 * test + 0{,}831 * gender$$
$$G1(0{,}95) = 3{,}70 + 0{,}023 * vbe + 0{,}429 * test + 0{,}686 * gender$$

As we can see all values have a trend to be the lowest in 0.05 quantile regression equation and highest to 0.95 quantile regression equation. Predictors have positive parameters. That means that females usually get higher final grade (as we mentioned above, gender is a pseudo variable with code 1 for female and 0 for male gender).

The regression coefficient estimates of independent variables across different quantiles compared to OLS estimation are shown in Fig. 1. The shaded area gives a confidence band of coefficients estimated across different quantiles. As seen from Fig.1, the influence of Intercept, VBE and gender is positive above 0.1 quantile. Intercept influence to the final grade increases steadily. Variable VBE reaches the strongest influence between 0.6 and 0.8 quantiles and decreases in the end. However, the influence of "test" significantly increases just in the beginning and the end. Variable "gender" gains the most stable influence through all quantiles.

The stepwise method for final grade specifies that variables are selected for the model based on a stepwise – regression algorithm. Multiple regression model for the final grade is:

$$G1 = 2{,}49139 + 0{,}02559 * vbe + 0{,}35609 * test + \\ + 0{,}64812 * gender$$

As we can see, all parameter estimates are positive. Most important variable is "test", for which partial R-Square is 59.25 percent. R-Square of this regression model is good – 69.65 percent. This means that 69.65 percent of variables are correctly predicted. Multicolinearity of model criteria was checked using variance inflation factor (VIF). The multiple regression model does not have multicolinear variables (all VIF results are below critical meaning of 4).

Residuals normality of linear regression model for final grade was calculated using the Shapiro-Wilk and Kolmogorow-Smirnov normality test. Both tests showed that residuals are normally distributed. The Breusch-Pagan and White tests showed that residuals are homoscedastic.

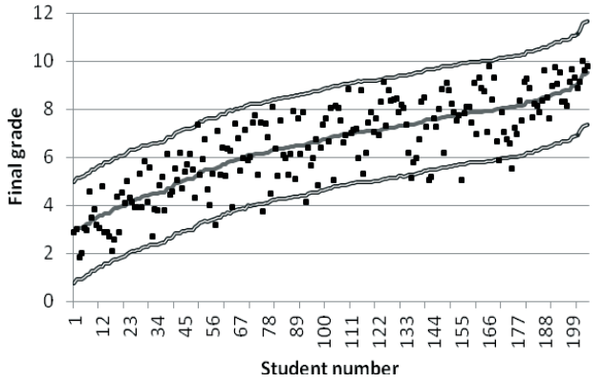

Fig. 2. Linear regression model with 95 percent Confidence Limits

In Fig. 2 we can see how the linear regression model (solid line) with 95 percent confidence limits (double lines) is situated compared to the final grades (dots).

Comparison of models

We have five quantile regression models and one linear regression model. For model comparison we choose to calculate the Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (8)$$

where A_t is the actual value and F_t is the forecast value. The absolute value in this calculation is summed for every fitted or forecasted point in time and divided again by the number of fitted points *n*. Multiplying by 100% makes it a percentage error. Table 1 below shows MAPE for final grade models:

Table 1. ***Mean Absolute Percentage Error (MAPE) for final grade models***

Model:	0,05	0,25	0,5	0,75	0,95	Linear
MAPE G1 %	30,24	16,35	14,88	20,32	29,96	14,98

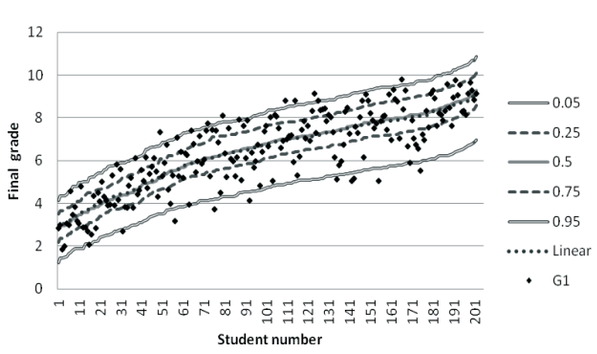

Fig. 3. Quantile and linear regression models for final grades

Median quantile regression model with 14.88 percent and linear regression model with 14.98 percent are the most accurate. Difference between them is very small. We can observe that MAPE is quite high for all models. The best decision, which can reduce MAPE, is to collect more data for research. Fig. 3 shows all models in one graph and how they are situated compared to final grades.

Conclusions

1. "Final grade" was forecasted using three directly unrelated criteria: national mathematics exam result (VBE), result of university mathematics test (test) and students' gender.
2. Model R-Squared determination coefficient showed that linear regression model accuracy is 69.95 percent. Therefore R-Squared indicates that the regression line fits the data very well.
3. Mean Absolute Percentage Error (MAPE) showed that for "final grade" forecast the most accurate are median quantile regression (14.88 percent error) and linear regression (14.98 percent error) models.
4. MAPE is quite high for all models; therefore, more data should be collected for more accurate results.

References

1. Buhai I. S., 2005, Quantile Regression: Overview and Selected Applications. *Ad Astra Journal.*
2. Cade B. S., Noon B. R., 2003, A gentle introduction to quantile regression for ecologists. *Front Ecol Eviron.* No. 1 (8). P. 412–420.
3. Collin L. C. *An Introduction to Quantile Regression and the QUANTREG Procedure*. SAS Institute.
4. Despa S., 2007, Quantile Regression. *StatNews.* No. 70.
5. Koenker R., 2013, *Quantile Regression in R: a vignette,* February 28, 2013.
6. Koenker R., Hallock K. F., 2001, Quantile Regression. *Journal of Economic Perspectives.* Vol. 15. No. 4. P. 143–156.
7. Mohr E., Dochow R., Schmidt G., 2012, October 10, A Quantile Regression Approach to Estimating Risk and Return. *Optimization*.
8. SAS customers support website. Available at: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#qreg_toc.htm [accessed 12-12-2013].
9. Seber G. A. F., Lee A. J. *Linear Regression Analysis. Second Edition.*

APPLICATION OF QUANTILE REGRESSION FOR THE PREDICTION OF STUDY RESULTS: THE CASE OF KTU

Laura Viselgaitė, Audrius Kabašinskas

Summary

In this research we analyze what influences the final grades of the Kaunas University of Technology module Mathematics 1.

The aim of the article is to find out how accurately we can predict the final grades of the Mathematics 1 module for each student in the beginning of semester, based on the national mathematics exam result and the pilot Kaunas University of Technology Mathematics test results. 217 students were chosen from the KTU faculty of Economics and Management, which participated in the module Mathematics 1 through 2009/10 and 2010/11 school years. To evaluate the final grade we chose these criteria: students' gender, national mathematics exam result and result of university mathematics test, which is taken by each student in the beginning of the semester. It was determined that national mathematics exam results and university mathematics test results are not normally (Gaussian) distributed. The final grade was predicted using models of quantile regression and a linear regression model was used for comparison. Quantile regression was made for 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles. All models were compared between each other and the mean absolute percentage error (MAPE) was evaluated. The median quantile regression and linear regression models were the most accurate.

Keywords: quantile regression, prediction, quantile, mathematics test.

KVANTILINĖS REGRESIJOS TAIKYMAS PROGNOZUOJANT REZULTATUS: KTU PAVYZDYS

Laura Viselgaitė, Audrius Kabašinskas

Santrauka

Darbe tiriama, nuo ko priklauso KTU Ekonomikos ir vadybos fakulteto pirmo kurso studentų galutinis modulio „Matematika 1" semestro įvertinimas. Tyrimo tikslas – nustatyti, kaip tiksliai semestro pradžioje galima prognozuoti galutinį kiekvieno studento modulio „Matematika 1" įvertinimą pagal valstybinio brandos egzamino ir Kauno technologijos universiteto bandomojo matematikos testo rezultatus.

Tyrimui buvo atrinkta 217 KTU Ekonomikos ir vadybos fakulteto studentų, kurie modulį „Matematika 1" mokėsi 2009–2010 ir 2010–2011 mokslo metais. Prognozuojant galutinį įvertinimą buvo remiamasi šiais kriterijais: studento lytis, matematikos valstybinio brandos egzamino (VBE) įvertinimas ir matematikos testo, laikomo mokslo metų pradžioje, pažymys. Nustatyta, kad studentų VBE ir testo įvertinimai nėra pasiskirstę pagal normalųjį (Gauso) dėsnį, todėl galutinis įvertinimas buvo prognozuojamas kvantilinės regresijos modeliais, o siekiant palyginti – ir tiesinės regresijos modeliu. Kvantilinei regresijai pasirinkti 0,05, 0,25, 0,5, 0,75 ir 0,95 kvantiliai. Visi gauti modeliai palyginti tarpusavyje ir apskaičiuota sudarytų modelių vidutinė absoliutinė procentinė paklaida (MAPE). Nustatyta, jog galutinis įvertinimas tiksliausiai prognozuojamas pagal medianinės kvantilinės regresijos ir tiesinės regresijos modelius.

Prasminiai žodžiai: kvantilinė regresija, prognozė, kvantilis, matematikos testas.

Įteikta 2014-01-12