Journal of Business Research 67 (2014) 779–784

Contents lists available at ScienceDirect

Journal of Business Research

A new quantile regression forecasting model ☆

Kun-Huang Huarng *, Tiffany Hui-Kuang Yu **

Feng Chia University, Taiwan

ARTICLE INFO

Article history:
Received 1 April 2013
Received in revised form 1 October 2013
Accepted 1 November 2013
Available online 13 December 2013

Keywords:
Health care expenditure
ICT
New quantile information criterion
Forecasting

ABSTRACT

Quantile regression is popular because it provides more information as well as comprehensive interpretations. To improve forecasting performance, this study proposes a new quantile information criterion (NQIC), on the basis of the coefficient of variation, and expects the NQIC to reflect whether a variable is predictable. The health care expenditure data determine the thresholds for the NQICs. The thresholds assist in forecasting the development of information and communication technology. From the empirical analyses, the NQICs and thresholds greatly improve the forecasting performance.

1. Introduction

The quantile regression model offers a more complete model than the conventional mean regression (Yu, Lu, & Stander, 2003). Studies apply the quantile regression model to interpret various problems, such as wages (Buchinsky, 1994; Machado & Mata, 2005; Martins & Pereira, 2004), survival analysis (Crowley & Hu, 1977; Koenker & Geling, 2001), financial analysis (Bassett & Chen, 2001), economic research (Hendricks & Koenker, 1992; Wang, Yu, & Liu, 2013), the study of the environment (Pandey & Nguyen, 1999), internet and communication technology (ICT) adoption (Yu, 2011), health care expenditure (Yu, Wang, & Chang, 2011), small business performance (Seo, Perry, Tomczyk, & Solomon, 2014), and so on.

A number of studies advance the quantile regression model to forecasting. Granger, White, and Kamstra (1989) propose a method for combining the variety of possible interval forecasts on the basis of quantile regression techniques. Taylor (2007) forecasts the daily supermarket sales using exponentially weighted quantile regression (interval forecasts), which outperforms traditional methods. Banachewicz and Lucas (2008) use hidden Markov models to forecast the quantiles of corporate default rates, which are important in financial risk management. Gerlacha, Chen, and Chan (2011) apply Markov chain Monte Carlo methods for the Bayesian time-varying quantile forecasting of Value-at-Risk in financial markets. Cai, Stander, and Davies (2012) propose a Bayesian approach to quantile autoregressive time series model estimation and forecasting and then apply the approach to currency exchange rate data. The empirical results show that an unequally weighted combining method outperforms other forecasting methodology.

Yu (forthcoming) suggests using a quantile information criterion (QIC) to assist in forecasting. To improve forecasting performance, this study proposes a new QIC (NQIC) to identify if a variable is predictable. The health care expenditure data are in order to determine the thresholds for the NQICs. Then, the thresholds and the NQICs intend to forecast the ICT development.

To that end, Section 2 reviews the concepts of the quantile regression model. Section 3 introduces the algorithms for the NQICs and determining their thresholds. Section 4 describes the variables of the two data sets, provides the empirical analyses for the two cases, and reveals the results of the estimation and forecasting performance. Section 5 concludes this paper.

2. Quantile regression model

Koenker and Bassett (1978) propose quantile regression to infer the results of the conditional functions for different quantiles. Bao, Lee, and Saltoğlu (2006) consider that the main advantage of quantile regression is to provide better statistics by means of the empirical quantiles. Quantile regression can help "complete the picture" when we intend to understand the relationship between variables for which the effects may vary with outcome levels. In addition, quantile regression is more forgiving than ordinary least squares in that quantile regression is

☆ The authors acknowledge and are grateful for the financial support provided by the National Science Council, Taiwan, ROC under grants NSC 101-2410-H-035-004-, NSC 100-2410-H-035-006-MY2, NSC 102-2410-H-035-038-MY2, and NSC 102-2410-H-035-041-MY2. The authors would also like to thank several attendees at the 2013 Conference of the Global Innovation and Knowledge Academy for their valuable comments that have helped improve this article.

* Correspondence to: K.-H. Huarng, Department of International Trade, Feng Chia University, 100 Wenhwa Road, Seatwen, Taichung 40724, Taiwan.
** Correspondence to: T. H.-K. Yu, Department of Public Finance, Feng Chia University, Taichung, Taiwan.
E-mail addresses: khhuarng@mail.fcu.edu.tw (K.-H. Huarng), hkyu@mail.fcu.edu.tw (T.H.-K. Yu).

0148-2963/$ – see front matter
http://dx.doi.org/10.1016/j.jbusres.2013.11.044

relatively insensitive to outliers and can avoid censoring problems (Conley & Galenson, 1998).

As Bassett and Koenker (1982) extend the median to quantile regression so as to calculate various quantiles, the quantile regression does not require any distribution assumptions regarding the population and can estimate the parameters nonparametrically. Quantile regression models the conditional quantiles, which are quantiles of the conditional distribution of the response variable in the expression of functions of the covariates of observations. Quantile regression models use the least absolute deviations method to minimize the absolute values of the errors. The model for a median linear regression is:

$$y_i = x_i\beta_\theta + \varepsilon_{\theta,i} \tag{1}$$

where the assumption is median $(\varepsilon_{\theta,i}|x_i) = 0$. This concept is extendable to any quantile, such as the 75th percentile, 95th percentile, etc. We can define the estimate by minimizing the sum of asymmetrically weighted residuals.

$$\min_{\beta}\left[\sum_{t|y_t \geq x_t\beta} \alpha|y_t - x_t\beta| + \sum_{t|y_t < x_t\beta} (1-\alpha)|y_t - x_t\beta|\right] \tag{2}$$

where α is a parameter $(0 < \alpha < 1)$ that represents the size of the quantile, and is also the quantile α of the explanatory variable that we intend to examine in the quantile regressions. This problem does not have an explicit form; however, linear programming methods can solve the problem (Armstrong, Frome, & Kung, 1979). When $\alpha = 0.5$, the quantile regression is the median regression. Since on this occasion the values of α and $(1 - \alpha)$ are both 0.5, the above equation changes to $\sum_y |y_t - x_t\beta|$, indicating that the observations above and below the median values are of the same weights.

3. New quantile information criterion

3.1. Rationale

Not all the variables are predictable by using quantile regressions due to their data characteristics. To identify if they are predictable, Yu (forthcoming) proposes a quantile information criterion (QIC). This study proposes a New QIC (*NQIC*) to provide a systematic method to improve the forecasting results.

To measure the dispersion of a distribution, this study applies the coefficient of variation as the NQIC, which is the ratio of the standard deviation and mean of a variable (Lind, Marchal, & Wathen, 2006). The coefficient of variation is unitless and is most useful in comparing the variability of different data sets (Rosner, 1995). For example, Bloch (2007) uses the coefficient of variation to test if the coefficients of two data sets are different. Yu (forthcoming) considers that the variables with large variations are difficult to forecast. Following these studies, this study first intends to identify the extreme values of the coefficients of variation of a data set to form the thresholds. Furthermore, the thresholds are in order to determine whether the variables in the other data set are unpredictable with the coefficients of variation lying outside the range of the thresholds.

3.2. Algorithms

There are two data sets: the one for obtaining the thresholds is the sample data set and the other is the target data set. Following the above rationale, this section proposes the algorithm for calculating the NQICs and their corresponding thresholds, and the algorithm for forecasting on the basis of the thresholds. We list the algorithms in Appendices 1 and 2.

Table 1
Quantile forecasting of 1998 by using 1992–1997 health care expenditure data.

1992–1997			1998
0.05			
LOG(GDP)	0.497261	1.390759	1.34671
OLD	0.013379	0.027361	0.01405
DOC	−0.012920	0.031001	0.03511
IM	−0.008170	−0.003110	−0.00234
LE	−0.019490	0.026669	−0.01954
0.25			
LOG(GDP)	1.056123	1.370417	1.12976
OLD	0.005542	0.015938	0.00192
DOC	0.025187	0.054073	0.04980
IM	−0.007550	−0.003070	−0.00262
LE	−0.017050	0.006608	0.01077
0.50			
LOG(GDP)	1.209432	1.577748	1.33037
OLD	0.001473	0.011047	0.01145
DOC	0.044355	0.088145	0.04925
IM	−0.009780	0.001800	−0.00216
LE	−0.02126	0.000844	−0.00336
0.75			
LOG(GDP)	1.391142	1.818118	1.52785
OLD	−0.006450	0.013389	0.00933
DOC	0.031638	0.091582	0.01192
IM	−0.003990	0.007852	0.00090
LE	−0.030450	−0.006490	−0.00700
0.95			
LOG(GDP)	1.579464	1.983476	1.92039
OLD	−0.041320	0.010996	−0.00434
DOC	0.005280	0.128820	0.07328
IM	−0.024830	0.010607	−0.00237
LE	−0.055570	−0.015770	−0.03289

The algorithm for calculating the *NQIC*s and thresholds serves to calculate the thresholds from the sample data set. Step 1 separates the data into in-sample and out-of-sample data. The calculation starts with the in-sample data ranging from $t = 1$ to $d - 1$ to test the out-of-sample data of $t = d$. The next run moves one step further; in other words, the calculation starts with the in-sample data from $t = 1$ to d to test the out-of-sample data of $t = d + 1$. We continue the process until we exhaust all the out-of-sample data.

Step 2 calculates the *NQIC*s and then checks if the corresponding quantile intervals can cover more than 50% of the corresponding variable in the out-of-sample data. Only those with more than 50% coverage (which we name *hits*) can advance to the calculation for the thresholds.

Step 3 finds the maximum (*max*) from all the positive *NQIC*s of all the hits and calculates their standard deviation (*stdev_p*). Similarly, it finds the minimum (*min*) from all the negative *NQIC*s and their standard deviation (*stdev_n*). To be conservative, we narrow the range between the minimum and maximum by one standard deviation, respectively.

In the algorithm for forecasting on the basis of the thresholds, we also separate the data into in-sample as well as out-of-sample data for the target data set in Step 1 of the previous algorithm. Step 2 calculates the *NQIC*s for the out-of-sample data.

During forecasting, if any $NQIC_t^v$ falls within the thresholds, the algorithm considers the variable v at time t predictable; otherwise

Table 2
The hits for all the years.

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
LOG(GDP)	**5**	**4**	**5**	**5**	**4**	**3**	**3**	**5**	2	**4**
OLD	**3**	**3**	**4**	**3**	**4**	**3**	**4**	**5**	**3**	**3**
DOC	**3**	**4**	**3**	**3**	**3**	**3**	**3**	**5**	0	1
IM	**3**	1	2	1	1	2	2	**5**	2	2
LE	**3**	**5**	**3**	2	**3**	**3**	**4**	**5**	2	**3**

Note: The numbers in bold are the hits.