

<https://doi.org/10.33472/AFJBS.6.9.2024.4693-4708>



African Journal of Biological Sciences

Journal homepage: <http://www.afjbs.com>



Research Paper

Open Access

A Study to Assess the Psychological Health of Young Adults through Quantile Regression and Quantile Regression Neural Network Models

Alka Sabharwal¹, Babita Goyal^{2,*} and Lalit Mohan Joshi³

¹ Professor, Department of Statistics, Kirori Mal College, University of Delhi, Delhi-91-110007, India

² Professor, Department of Statistics, Ramjas College, University of Delhi, Delhi-91-110007, India

³ Research scholar, Department of Statistics, University of Delhi, Delhi-91-110007, India

* Correspondence:

Babita Goyal, Professor, Department of Statistics, Ramjas College, University of Delhi, Delhi-91-110007, India.

goyalbabita@gmail.com

Article Info Volume 6, Issue 9, 2024

Received: 09 Apr 2024

Accepted: 10 May 2024

doi:10.33472/AFJBS.6.9.2024.4693-4708

Abstract

For qualitative and skewed data collected through questionnaires, which do not meet the assumptions of classical regression models, response variables are estimated through alternate regression models. With the objective to obtain significant predictors on the psychological health of young adults, Quantile regression (QR) and Quantile regression Neural Network (QRNN) models were employed on the data collected through three chronological surveys on young adults in higher educational institutions in India during COVID-19 period using the Strength and Difficulty Questionnaire.

The QR model was applied with the quantile obtained as ratio of frequencies of two extreme categories. For the QRNN model, optimal quantiles were extracted from a range of quantiles on the basis of the maximum predictive power. The two models were compared through mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). The efficacy of the models was computed and predictive powers of both the models were obtained under each category of the response variable by comparing the results with the observed data. Significant predictors were extracted through the QR model. Emotional symptom and hyperactivity-inattention problem were found to be the significant predictors in all the three surveys.

On the basis of predictive power, QRNN performed better than the QR models. But the significant predictor can't be extracted through QRNN as the model does not provide the significant value of the test statistic. The study concluded that the two models should be used simultaneously to get a comprehensive picture of data under study.

Keywords: Behavioral problems, Psychological health, Quantile Regression, Quantile Regression Neural Network, Social dysfunction

1 Introduction

Generally, the regression models are used to estimate the conditional mean of a response variable for a given set of predictors. Among various regression models, the most frequently used models are multiple regression, neural networks, support vector machine (SVM), decision trees, etc. , each used in a specific scenario but with the same objective of estimating the mean of the response variable. However, the situation may not always warrant the prediction about the mean. If the data is skewed, one may be more interested in the use of different quantiles rather than the mean only. But the use of quantiles increases the complexity of the model and the number of calculations. However, advancements in computing have allowed the development of regression models for predicting a given quantile of the conditional distribution, both parametrically and non-parametrically (Vasseur *et al.*, 2021). The general approach is called **Quantile Regression**, but the methodology (of conditional quantile estimation) applies to any statistical model, be it multiple regression, SVM vector machines, or random forests. Quantile regression neural networks (QRNN) estimate conditional values of an individual quantile using a multilayer perceptron neural network (Taylor, 2000).

Quantile regression (QR), introduced and developed by Koenker and Bassett in 1978, complements and improves the traditional mean regression models which assume that the data is distributed homogeneously. However, this assumption is rarely met and in general data sets tend to be asymmetric. In such cases when asymmetries and heavy tails exist, the sample median, the 50th percentile as quantile, provides a better summary of centrality than the mean. QR, which estimates the unknown parameters using various quantiles, is more robust to outliers and more flexible as in QR the outcome does not need to be strictly specified in the parametric distribution assumptions. This approach assumes the various quantile functions of a Y distribution as a function of X (Haug *et al.*, 2017; Koenker *et al.*, 2005; Koenker, 2005).

Mostly quantile regression is used for prediction with multiple predictors based on linear or simple parametric nonlinear models (Koenker *et al.*, 2005). A further advancement in QR was suggested by Taylor (2000) when he proposed Quantile Regression Neural Network (QRNN), a nonlinear nonparametric method that combined the advantages of quantile regression and neural network (Taylor, 2000). Neural Network (NN) is a forecasting method applied to nonlinear data with high accuracy, without requiring a principle to meet (Zhang, 2003). Hidden layers are inherent in NN which are nodes, each meant to perform a specific, predetermined function and to produce some output, which then becomes the input of the second node, if any. For prediction, a neural network with a single hidden layer feed-forward network which consists of a set of inputs is widely used.

Due to its flexible and versatile nature, the QR has been applied to very different types of areas viz. investment, finance, economics, medicine, and engineering. Huang *et al* (2017) applied QR to different types of data; independent data, time-to-event data, and longitudinal data (Haug *et al.*, 2017). Telbany (2014) applied the QRNN technique to predict the activity of the pyrimidines-based structure–activity relationship of these compounds for finding potential treatment agents for serious diseases. The study also compared QR and QRNN models and showed that the QRNN model significantly reduced the prediction error (El-Telbany, 2014).LI Gan-qiong *et al.*(2012)used the approach to model

price movements of livestock based food items and to analyse the short-term impacts of cost factors on livestock's price fluctuation. They also generated confidence intervals from quantiles 5th and 95th and showed that it is a good method to forecast livestock price fluctuation (Liet *et al.*, 2012). Sigalingginget *et al.*, (2004) and Amalia *et al.*, (2018) applied QRNN to forecast inflow and outflow of cash (Sigalingging *et al.*, 2004; Amalia *et al.*, 2018). In the present study, the quantile regression model (QR) and quantile regression neural network (QRNN) models have been used to estimate/predict the categorical response variable with independent predictors. Both models were applied to the real data collected during the COVID-19 pandemic (Goyal *et al.*, 2023) with the help of the Strength and Difficulty Questionnaire (SDQ) 17⁺ extended version concerning the psychological health of young adults in India. Three surveys were conducted at different time points; and, respectively 1020, 743, and 934 responses were obtained. Among the two components of the extended version, the 'impact' score was the response variable, which measures distress in terms of social dysfunction, and the scales of the second component viz. the 'difficulty' score were taken as the predictors that measure behavioral problems. These scales are the conduct problem, peer problem, emotional symptoms, and hyperactivity-inattention. Firstly, the QR model was applied with the appropriate value of the quantile, and significant predictors were extracted. Secondly, for the application of the QRNN model, the data was divided randomly into two parts; train data and test data. The QRNN model was applied with the train data for the range of quantiles with one hidden node, and the regression coefficients were computed for every value of quantile. Further, for each set of regression coefficients, the response variable was estimated for the test data. Thirdly, the QR and the QRNN models were compared with mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). Lastly, we estimated the response variable values/category from the predictors under both models, and the efficacy of the models was computed and compared. The predictive power of both models was computed by comparing the results with the observed values.

1.1 Objectives and Novelty of the study

This study is in continuation with a previous work by authors in which they used Ordinal regression to estimate the three severity categories of the impact score with the components of the difficulty score. Although the model was found to be efficient for estimating the categories of the 'Normal' and the 'Abnormal' cases but the model failed to estimate the 'Borderline' category cases (Sabharwal *et al.*, 2023). With the objective of overcoming the limitation of the OR model for this complex and multi-component data, and to get a complete picture in terms of the significant predictors and optimal quantile the authors explored the use of QR and QRNN models.

Novelty of this study is that the applied models have been cross-validated through three different data sets in place of cross-validation through the same data set which is the general procedure followed in machine learning. Also the robustness of the QR models has been established by comparing standard median-based models with data-specific quantiles. To the best of our knowledge, this is the first study where QR and QRNN models were compared after extracting the optimum value of the quantile. Further, this comparison helped us to extract significant predictors under QRNN models, which was otherwise not possible from its result table.

Besides the introduction, the course of the paper is as follows: material and methods have been explained in Section 2; applications of the models are discussed with real data under results in Section 3; followed by discussion and conclusion in Section 4.

2 Material and Methods

2.1 Material

The data collected and the method of data collection are therein the previous works of the authors (Goyal *et al.*, 2023; Sabharwal *et al.*, 2023) as the present study is in continuation with these works.

2.2 Methods

2.2.1 Quantile Regression

Quantile Regression uses an asymmetric weighting system of data which are based on their distance from the data-specific quantile for the estimation (Farooqi, 2019). A multiple Quantile Regression model with k predictors can be written as

$$Y = \beta_{qk} X + \varepsilon \quad (2.2.1)$$

where Y is $(n \times 1)$ matrix of the response variable, X is an $(n \times (k+1))$ matrix of the predictor variables (including the intercept), β_{qk} is a $(k \times 1)$ matrix of the regression coefficients at given quantile level and ε is a $(n \times 1)$ matrix of the error.

For a single predictor X_i , model (1) is reduced to

$$Y_i = \beta_{q0} + \beta_{qi} X_{qi} + \varepsilon_{qi} \quad (2.2.2)$$

where ε_{qi} 's are random error, β_{q0} and β_{qi} are the unknown parameters associated with q^{th} quantile.

The q^{th} quantile given X_i is defined as:

$$Q_q(Y_i | X_i) = \beta_{q0} + \beta_{qi} X_{qi} \quad (2.2.3)$$

$$Q_q(\varepsilon_{qi}) = 0$$

Let $\hat{\beta}_{q0}$ and $\hat{\beta}_{qi}$ be the estimates of the unknown parameter β_{q0} and β_{qi} respectively.

To obtain the estimates, the sum of absolute errors, $\sum_{i=1}^n |\varepsilon_{qi}|$ is minimized subject to asymmetric penalties $q(0 < q < 1)$ for under prediction and $(1-q)$ for over prediction (Farooqi, 2019) so that the objective function to be minimized is

$$Q(\beta_q) = \min \left[q \sum_{i:Y_i \geq X_{i\beta}} |\varepsilon_{qi}| + (1-q) \sum_{i:Y_i < X_{i\beta}} |\varepsilon_{qi}| \right]$$

where,

$$\varepsilon_{qi} = Y_i - (\beta_{q0} + \beta_{qi} X_{qi})$$

So,

$$\begin{aligned} Q(\beta_q) &= \min \left[q \sum_{i:Y_i \geq X_{i\beta}} |Y_i - (\beta_{q0} + \beta_{qi} X_{qi})| + (1-q) \sum_{i:Y_i < X_{i\beta}} |Y_i - (\beta_{q0} + \beta_{qi} X_{qi})| \right] \\ &= \min \left[\sum_{i=1}^n I_{(Y_i \geq X_{i\beta})} q |Y_i - (\beta_{q0} + \beta_{qi} X_{qi})| + (1-q) I_{(Y_i < X_{i\beta})} |Y_i - (\beta_{q0} + \beta_{qi} X_{qi})| \right] \end{aligned} \quad (2.2.4)$$

where I is an indicator function.

2.2.2 Goodness of Fit test for Quantile Regression

A measure of goodness of fit is obtained by comparing the sum of the weighted absolute deviations of the full model (a model with all the predictors in it) with the, that of the null model (no predictors in the model)(Koenker *et al.*, 1999). So if $\hat{Z}(q)$ is the sum of the weighted absolute deviations of the full model and $\tilde{Z}(q)$ that of the null model for prediction of the same quantile q then a measure of goodness of fit for quantile regression is:

$$R(q) = 1 - \frac{\hat{Z}(q)}{\tilde{Z}(q)}; 0 \leq R(q) \leq 1 \tag{2.2.5}$$

The larger value of $R(q)$ is indicating a better model fit.

The likelihood ratio (LR) test measures the difference in the absolute deviations of null model and the full model as a proportion of null model at the q^{th} quantile (Koenker *et al.*, 1999).i.e.

$$LR = \frac{2(\tilde{Z}(q) - \hat{Z}(q))}{q(1-q)\{f_{eq}(0)\}^{-1}} \sim \chi^2_{(k)} \tag{2.2.6}$$

Where $\{f_{eq}(0)\}^{-1}$, the reciprocal of the density function evaluated at the quantile of interest, is called the sparsity function; and is considered a nuisance parameter (Koenker *et al.*, 1999).

$$\tilde{Z}(q) = \sum_{i=1}^n |Y_i - \beta_0|; \text{ and}$$

$$\hat{Z}(q) = \sum_{i=1}^n |Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ki})|;$$

and k is the difference between the number of predictors in full model and null model.

2.2.3Quantile Regression Neural Network (QRNN)

A QRNN model works as follows:

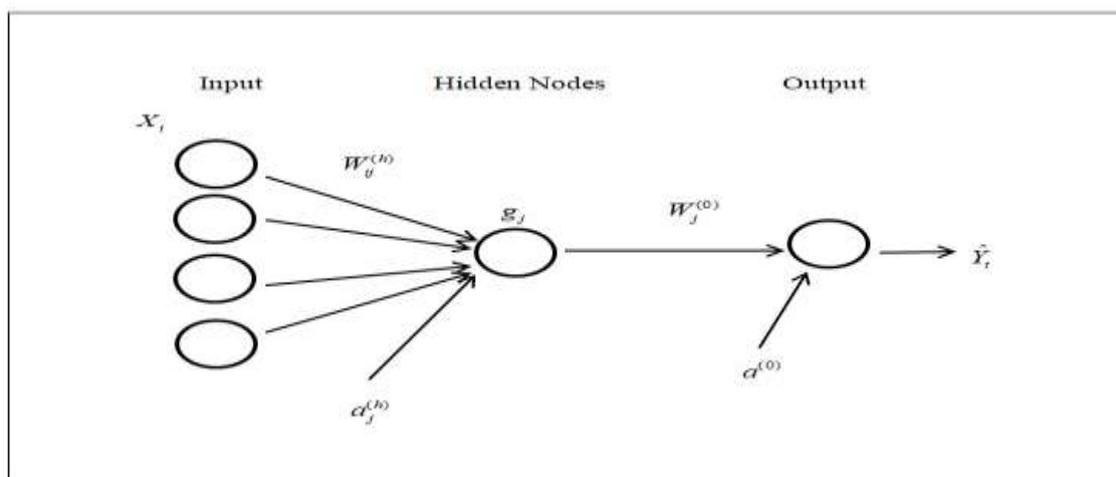


Figure1:Quantile Regression Neural Network model with four predictors and with one hidden layer

For given predictors, $X_i(t)$ and response variable, $Y(t)$, the output of at a hidden layer node j is,

$$o_i(t) = \tanh\left(\sum_{i=1}^I X_i(t)W_{ij}^{(h)} + a_j^{(h)}\right) \tag{2.2.7}$$

where $W_{ij}^{(h)}$, $a_j^{(h)}$ are the weight, and bias of hidden layer respectively. An estimate of $\hat{Y}_q(t)$, the conditional q^{th} quantile is,

$$\hat{Y}_q(t) = f\left(\sum_{j=1}^J o_i(t) + W_j^{(0)} + a^{(0)}\right) \quad (2.2.8)$$

where $f(\cdot)$ is a function of the output layer with $W_j^{(0)}$ and bias $a^{(0)}$ at the output layer (Cannon,2011).

To minimize the loss between the predicted and target outputs, a **loss function** is used to **compare** the target and predicted output values of the train data. For QRNN, **Huber Loss** is used generally, which has the advantage over MSE and MAE when the average distance approaches 0 and the function's derivative at 0 are undefined.

The Huber Loss function is defined as

$$\rho_q(e) = \begin{cases} q h(e) & \text{if } e \geq 1 \\ (q-1)h(e) & \text{if } e < 0 \end{cases} \quad (2.2.9)$$

where e the residual of the parameter and $h(e)$ is Huber norm given by

$$h(e) = \begin{cases} \frac{e^2}{2\varepsilon} & \text{if } 0 \leq |e| \leq \varepsilon \\ |e| - \frac{\varepsilon}{2} & \text{if } |e| \geq \varepsilon \end{cases} \quad (2.2.10)$$

ε is the threshold whose value is to be determined. Then the optimized error function is

$$E_q = \frac{1}{N} \sum_{t=1}^N \rho_q\left(Y(t) - \hat{Y}_q(t)\right) \quad (2.2.11)$$

2.2.4 Comparison of QR and QRNN Models

QR and QRNN models are compared using Mean absolute Error (MAE), Mean square Error (MSE) and Root mean Square Error (RMSE).

$$MAE = \frac{1}{m} \sum_{i=1}^m |Y_i - \hat{Y}_i| \quad (2.2.12)$$

$$MSE = \frac{1}{m} \sum_{i=1}^m (Y_i - \hat{Y}_i)^2 \quad (2.2.13)$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (Y_i - \hat{Y}_i)^2} \quad (2.2.14)$$

\hat{Y}_i are the predicted values and Y_i , the observed values.

3 Results

Cronbach's alpha and Guttman Lambda (Goyal *et al.*, 2023) coefficients of data reliability were respectively 0.83 and 0.89 (for first survey), 0.86 and 0.89 (for second survey); and 0.88 and 0.86 for third survey. Table 1 has the descriptive statistics for all five scales of the basic version and the impact score of the extended version component. The descriptives have been presented only for those respondents who were facing distress due to behavioural problem(s). Impact scores were available for: 772/1020, 584/743, and 773/934 from Survey 1, Survey 2, and Survey 3 responses respectively.

Table1: Descriptive statistics of all three surveys giving mean, standard deviation, minimum, and maximum of five strength and difficulty scales; and Difficulty and Impact scores

Scale (Items)	Surveys	Total	Mean	Sd	Minimum	Maximum
Prosocial behaviour (1, 4, 9, 17, 20)	1	772	7.891	1.686	1	10
	2	584	7.932	1.754	0	10
	3	773	7.825	1.736	0	10
Hyperactivity-inattention (2, 10, 15, 21, 25)	1	772	4.104	2.034	0	9
	2	584	3.724	2.059	1	10
	3	773	4.304	1.916	0	10
Emotional symptoms (3, 8, 13, 16, 24)	1	772	4.193	2.450	1	10
	2	584	3.995	2.488	0	10
	3	773	4.788	2.519	0	10
Conduct problem (5, 7, 12, 18, 22)	1	772	2.935	1.456	0	9
	2	584	2.785	1.441	1	7
	3	773	3.119	1.554	0	8
Peer problem (6, 11, 14, 19, 23)	1	772	2.902	1.713	0	10
	2	584	2.942	1.787	0	9
	3	773	3.352	1.884	0	10
Difficulty score	1	772	14.136	5.142	1	31
	2	584	13.443	5.568	2	33
	3	773	15.562	5.499	2	31
Impact Score (28, 29,30,31,32)	1	772	1.528	1.721	0	7
	2	584	1.885	2.100	0	9
	3	773	2.289	2.558	0	10

1.2 The QR model

1.2.1 Testing the validity of the model

For every participant, the QR model with independent predictors: conduct problem, peer problem, emotional symptoms; and hyperactivity-inattention, has been applied to estimate the probability/category of impact score. The data based quantile q is taken to be the 'ratio of problematic to normal category' for all the three surveys; yielding values 0.44, 0.36, and 0.34 for survey 1, survey 2, and survey 3 respectively. Using the computed value of q , the coefficient of determination, $R(q)$ is computed for all the three surveys, which in turn is used in the likelihood ratio (LR) test to compare the null model with the full model with given predictors.

Null hypothesis and alternative were:

H_{0i} : There is no significant difference between the null and full model.

H_{1i} : There is significant difference between the null and full model.

i = survey 1, survey 2, survey 3

The results are presented in Table 2 below:

Table 2: - Likelihood Ratio test for all three surveys data

Survey	q	$f_{eq}(0)$	χ^2 -statistic	p -value
1	0.44	0.22	397.023	<0.0001
2	0.36	0.18	431.835	<0.0001
3	0.34	0.17	656.080	<0.0001

With each p -value less than 0.001, the full model is significantly different from the null model for all the three surveys.

Concerning the significance of individual predictors, hyperactivity-inattention, and emotional symptoms are significant contributors to estimating respondents' probability of belonging to a category, as $p < 0.01$ in all three surveys. Conduct problem is a significant factor in survey 1. Peer problem is a significant contributor for estimating impact score only in survey 2. The detailed results are given below in Table 3.

Table3: Quantile Regression showing the partial effects of components of difficulty scales on impact scores of the participants in the three surveys

First Survey $q=0.44$ overall				
Coefficients	Estimate	Std. Error	t- value	Pr(> t)
Intercept	-0.57407	0.14361	-3.99742	0.00007
Hyperactivity Scale	0.16667	0.03231	5.15896	0.00000
Emotional	0.22222	0.03678	6.04225	0.00000
Peer Problem	0.05556	0.03781	1.46935	0.14215
Conduct Problem	-0.09259	0.03980	-2.32659	0.02025
Fitted Model				
Q((Impact components of difficulty score)=-0.57407+0.16667*Hyper+0.2222*Emo+.055556*Peer-0.09259259*Con				
P(Impact c<-0.57407+0.16667*Hyper+0.2222*Emo+.055556*Peer-0.09259259*Con)=0.44				
Second Survey $q=0.36$ overall				
Coefficients	Estimate	Std. Error	t- value	Pr(> t)
Intercept	-0.87500	0.19001	-4.60508	0.00001
Hyperactivity Scale	0.12500	0.03839	3.25622	0.00120
Emotional	0.25000	0.03733	6.69747	0.00000
Peer Problem	0.12500	0.04177	2.99225	0.00289
Conduct Problem	0.00000	0.04902	0.00000	1.00000
Fitted Model				
Q((Impact components of difficulty score)=-0.875+0.125*Hyper+0.25*Emo+.125*Peer-0.00*Con				
P(Impact c<-0.875+0.125*Hyper+0.25*Emo+.125*Peer-0.00*Con)=0.36				
Third Survey $q=0.34$ overall				
Coefficients	Estimate	Std. Error	t- value	Pr(> t)
Intercept	-1.14119	0.21854	-5.22196	0.00000
Hyperactivity Scale	0.14936	0.03144	4.75030	0.00000
Emotional	0.24737	0.03879	6.37693	0.00000
Peer Problem	0.09568	0.04942	1.93615	0.05322
Conduct Problem	0.03734	0.06137	0.60847	0.54306
Fitted Model				
Q((Impact components of difficulty score)=-1.14119+0.14936*Hyper+0.24737*Emo+.09568*Peer-0.03754*Con				
P(Impact c< -1.14119+0.14936*Hyper+0.24737*Emo+.09568*Peer-0.03754*Con)=0.34				

1.2.2 Quantile regression Neural Network (QRNN) model

To apply QRNN, the total data was divided into two parts; train data, which was 60% of the total data, selected by random sampling and the remaining 40% as the test data. The quantiles were computed for the train data yielding values 0.43 for survey 1, 0.40 for survey 2, and 0.36 for survey 3. For each survey a narrow band of quantiles was selected which included the computed value in it and QRNN with a single hidden node was run for each quantile in the chosen bands. The results are given below in Table 4.

Table 4: Estimated value of predictor coefficient for different values of q for QRNN model for train data with one hidden node

First Survey					
Values of q	0.40	0.42	0.46	0.48	0.43
Coefficients	Estimated value				
Intercept	-0.50710	-0.46178	-0.35809	-0.27098	-0.40394
Hyperactivity Scale	0.13333	0.13334	0.13334	0.13334	0.13334
Emotional	0.26666	0.26665	0.26666	0.26666	0.26665
Peer Problem	0.01512	0.01511	0.01512	0.01509	0.01509
Conduct Problem	-0.10256	-0.10255	-0.10254	-0.10254	-0.10254
Second Survey					
Values of q	0.32	0.34	0.38	0.40	
Coefficients	Estimated value	Estimated value	Estimated value	Estimated value	
Intercept	-1.13780	-1.13780	-1.13780	-1.13780	
Hyperactivity Scale	0.14754	0.14754	0.14754	0.14754	
Emotional	0.30491	0.30491	0.30491	0.30491	
Peer Problem	0.12457	0.12457	0.12457	0.12457	
Conduct Problem	0.04912	0.04912	0.04912	0.04912	
Third Survey					
Values of q	0.30	0.32	0.36	0.38	
Coefficients	Estimated value	Estimated value	Estimated value	Estimated value	
Intercept	-0.929835	-0.877123	-0.754416	-0.684285	
Hyperactivity Scale	0.105263	0.105283	0.105279	0.105280	
Emotional	0.228076	0.228065	0.228075	0.228073	
Peer Problem	0.087722	0.087713	0.087708	0.087712	
Conduct Problem	-0.017543	-0.017571	-0.017550	-0.017558	

Table 5 below presents the comparison of the QRNN model with the QR model using mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). For all the three surveys, errors were lower under the QRNN model than the errors under QR model.

Table 5: Comparison of QR and QRNN models based on MAE, MSE and RMSE for different values of q

First Survey								
QR for q		QRNN for q						
	0.44	0.50	0.44	0.40	0.42	0.46	0.48	0.43
MAE	0.7261	0.9186	0.7273	0.7263	0.7306	0.7366	0.7523	0.7227
MSE	0.8552	1.2491	0.8455	0.8635	0.8770	0.8701	0.8792	0.8461
RMSE	0.9247	1.1176	0.9195	0.9292	0.9365	0.9328	0.9376	0.9198
Second Survey								
QR for q		QRNN for q						
	0.36	0.50	0.36	0.32	0.34	0.38	0.40	

MAE	0.6949	1.2525	0.6912	0.7049	0.7035	0.6948	0.6801
MSE	0.8174	2.2330	0.8174	0.8535	0.8629	0.8128	0.8086
RMSE	0.9041	1.4943	0.9011	0.9238	0.9289	0.9015	0.8992
Third Survey							
	QR for q		QRNN for q				
	0.34	0.50	0.34	0.30	0.32	0.36	0.38
MAE	0.6761	1.1130	0.6758	0.7081	0.6933	0.6627	0.6794
MSE	0.7007	1.9840	0.7007	0.7600	0.7346	0.6974	0.7279
RMSE	0.8371	1.4085	0.8370	0.8717	0.8571	0.8351	0.8532

In order to select an optimum QRNN model, results obtained using one; two and three hidden nodes for the optimum quantile (the one with minimum error term) were compared on the basis of MAE, MSE, and RMSE values. The results in Table 6 show that a single hidden node was the best choice for QRNN model for all the three surveys.

Table 6: Justification of hidden nodes for all three surveys by using MAE, MSE, and RMSE

First Survey			
Hidden nodes	1	2	3
MAE	0.7281	0.7587	0.7551
MSE	0.8484	0.9296	0.9122
RMSE	0.9210	0.9641	0.9551
Second Survey			
MAE	0.6941	1.0275	1.0317
MSE	0.8162	1.5180	1.5391
RMSE	0.9034	1.2320	1.2406
Third Survey			
MAE	0.6758	0.7048	0.7236
MSE	0.7007	0.7741	0.8064
RMSE	0.8370	0.8798	0.8980

The regression results of QR and QRNN models (corresponding to the optimum quantile for each survey) are presented in Table 7 below. A comparison of the results of the two regressions shows that the two models yielded almost similar estimates of the independent predictors. This finding helped us to overcome a limitation of QRNN model which does not provide the significance of individual predictors. With $p < 0.05$ in all three surveys, the hyperactivity-inattention and emotional symptoms were found to be significant contributors to the probability that a respondent will belong to a category. For the other two predictors, conduct problem was significant survey 1; and peer problem in survey 2 only.

Table 7: Comparison of QR and QRNN predictor coefficient for the computed quantile

First Survey	
QR	QRNN

Coefficients	Estimate value	Pr(> t)	Estimate value
Intercept	-0.32658	0.0979	-0.40394
Hyperactivity Scale	0.13165	0.0012	0.13334
Emotional	0.26582	0.0000	0.26665
Peer Problem	0.01013	0.8606	0.01509
Conduct Problem	-0.11899	0.0245	-0.10254
Second Survey			
Intercept	-1.1019	0.00001	-1.13780
Hyperactivity Scale	0.1337	0.00304	0.14754
Emotional	0.3152	0.00000	0.30491
Peer Problem	0.1242	0.02284	0.12457
Conduct Problem	0.0350	0.60369	0.04912
Third Survey			
Intercept	-0.7647	0.00652	-0.82452
Hyperactivity Scale	0.0980	0.03037	0.10527
Emotional	0.2156	0.00005	0.22806
Peer Problem	0.0980	0.08953	0.08770
Conduct Problem	-0.0198	0.76189	-0.01755

Table 8 shows the expected frequency of each category of impact score corresponding to different categories of difficulty scores was computed using both the models.

Table 8: Comparison of observed values with estimated values under both the models for all the three survey

First Survey				
Difficulty		Impact Observed	QR Estimated (q=0.44)	QRNN Estimated (q=0.43)
Normal (489)	Normal	283	231	256
	Borderline	34	239	224
	Abnormal	172	19	09
Borderline (168)	Normal	46	05	05
	Borderline	28	78	104
	Abnormal	94	85	59
Abnormal (115)	Normal	14	00	00
	Borderline	24	17	30
	Abnormal	77	98	85
Second Survey				
Difficulty		Impact Observed	QR Estimated (q=0.36)	QRNN Estimated (q=0.40)
Normal (391)	Normal	190	185	191
	Borderline	73	190	148
	Abnormal	128	16	52
Borderline (106)	Normal	19	00	05
	Borderline	21	32	55
	Abnormal	66	74	46
Abnormal (87)	Normal	05	00	08
	Borderline	09	01	13
	Abnormal	73	86	66
Third Survey				
Difficulty		Impact Observed	QR Estimated (q=0.34)	QRNN Estimated (q=0.36)
Normal (335)	Normal	189	203	190
	Borderline	66	130	136
	Abnormal	80	02	09
Borderline (293)	Normal	65	02	07
	Borderline	42	141	175
	Abnormal	186	150	111
Abnormal (145)	Normal	15	00	00
	Borderline	18	01	08
	Abnormal	112	144	137

While the QR model estimated the observed 'Normal' band as the 'Normal' and the observed 'Advance' band as the 'Advance' with over 83% accuracy, the QRNN model predicted the same with over 88% accuracy. The participants with overall difficulty score under the 'Normal' band but facing 'a great deal' problem in one of the areas of behavioural problems were estimated in the 'Borderline' category of the impact score. The borderline cases were predicted with 40 % accuracy. Further, the predictive power of QRNN model is better than QR model for all the three surveys as shown in Figure 2 below.

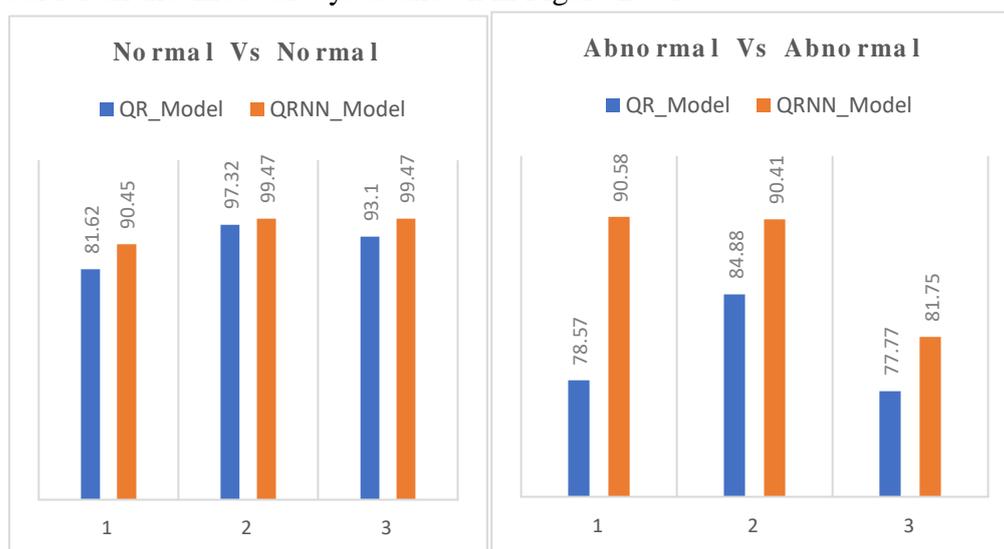


Figure 2: The predictive power (%) of the two models for the 'normal' and 'abnormal' categories of impact scores

All the analysis has been done using R software, version 4.2.1. The R-code is available in the Appendix.

2. Discussion and conclusion

In an earlier work, the authors used the ordinal regression (OR) model to estimate social dysfunction with components of difficulty scores (Sabharwal *et al.*, 2023). While the OR model exhibited good predictive power to estimate the normal and abnormal categories of social dysfunction as measured by impact score, it was not able to estimate the borderline category. The idea of this study was to overcome this limitation of the applied ordinal regression. One of the possible solutions was exploring the use of QR and QRNN models as the data was skewed. Both QR and QRNN models have exhibited good predictive powers as the models are based on quantile value. QR provides greater flexibility than other regression methods to identify differing relationships at different parts of the distribution of the dependent variable (Cannon, 2011). Further these models are not just based on median but the quantile(s) can be researchers' choice based thus increasing the flexibility of the model(s). Here, we are interested in group-differences across the distribution of a given dependent variable rather than only the mean. We used quantile regression to allow slopes of the regression line to vary across quantiles of the psychological health scale. The median-based quantile prediction was compared with data-based quantile prediction values and found to be significantly different. However, under the QR model, one can use only one quantile value, i.e. for comparing the model for different values of quantiles one has to run the model the same number of times as the number of quantile values. But the QRNN model is one-step

ahead of the QR model. Under QRNN we have used a range of quantile values and checked the predictive power for each quantile value, thus suggesting the most appropriate quantile value and corresponding coefficients of independent predictors for the train data. Using the obtained coefficients, we predicted the impact score/category for the remaining data (40%). However QRNN gives the optimum value of the quantile, based on the highest predictive power but the significant predictor can't be extracted as the test statistic along with that the significant value is not there for all predictors. This justifies the use of QR and QRNN models together to get appropriate quantile and the effect of predictors on the dependent variable based on that chosen quantile.

The comparison of frequencies obtained through QRNN model and those through the QR model (Table 8) yielded that QRNN was better in estimating/predicting distress categories as compared to QR. These findings were consistent with the observations of (EL-Telbany, 2014). As depicted in Figure 2, the predictive power of the models was found to be good as the categories of respondents having no distress or severe distress were estimated/predicted with 90% accuracy by both the models which is similar to the results produced by the OR model. In spite of the complex and multi-component nature of the SDQ questionnaire, both these models were able to estimate/predict the borderline category with 40% accuracy as against no prediction by the OR model (Sabharwal *et al.*, 2023).

The models were applied to the survey-based data (as against the simulated data which has been the usual practice). The study not only applied the models but also validated the use of the models for psychological data as the models were applied on three distinct data sets obtained at different time points during COVID-19 period and each time consistent results were obtained.

For data based quantile models, the upper bound of predictive power lies at the optimum quantile obtained. The classical confidence intervals are meaningless for these models.

This study was a step further than the previous study based on OR model but this is not the last model for this SDQ data as the authors wish to explore some other models also. A limitation of this study is that the models have not been explored for longitudinal data.

References

- Amalia FF, Rahayu SP, and Suhermi N (2018, June). Quantile regression neural network for forecasting inflow and outflow in Yogyakarta. *In Journal of Physics: Conference Series* (Vol. 1028, No. 1, p. 012232). IOP Publishing.
- Cannon AJ (2011). Quantile regression neural networks: implementation in R and application to precipitation downscaling. *Computers & Geosciences*, **37**: 1277-1284. doi:10.1016/j.cageo.2010.07.005
- El-Telbany ME (2014). Quantile Regression Neural Networks Based Prediction of Drug Activities.
- Farooqi A (2019). A Comparative Study of Kendall-TheilSen, Siegel Vs Quantile Regression with Outliers (2019). *Wayne State University Dissertations*. 2352.
- Goyal B, Sabharwal A, Chauhan V, & Joshi, LM (2023). Psychological health of Indian youth during COVID-19: a study through three chronological surveys. *International Journal of Public Health*, **12(2)**, 752-763.
- Huang Q, Zhang H, Chen J, & He MJBB(2017). Quantile regression models and their applications: a review. *Journal of Biometrics & Biostatistics*, **8(3)**, 1-6.

- Koenker R (2005). Quantile Regression. In: Econometric Society Monographs, No. 38. Cambridge University Press, Cambridge, New York 368pp.
- Koenker R. and Machdo JAF (1999). Goodness of fit and related inferences processes for Quantile Regression: *Journal of the American Statistical Association*. **94**:1296-1310.<http://dx.doi.org/10.1080/01621459.1999.10473882>.
- Koenker RW & Bassett GW(2005). Qunatile Regression, Cambridge: *Cambridge University Press*.
- LI, GQ, XU SW, LI Z M, SUN YG, &DONG XX (2012). Using quantile regression approach to analyze price movements of agricultural products in china. *Journal of Integrative Agriculture*, 11(4), 674-683.
- Sabharwal A, Goyal B, & Joshi LM (2023). An application of ordinal regression to extract social dysfunction levels through behavioral problems. *AIMS Public Health*, **10(3)**: 577-592.
- Sigalingging H, Setiawan E, Sihaloho HD (2004). KebijakanPenedaranUang Di Indonesia Jakarta: PusatPendidikandanStudiKebanksentralan.
- Taylor JW (2000). A Quantile Regression Neural Network Approach to Estimating the Conditional Density of Multiperiod Returns. *Journal of Forecasting*, 19 299-311.
- Vasseur SP, AznarteJL(2021). Comparing quantile regression methods for probabilistic forecasting of NO₂ pollution levels. *Sci Rep* **11**, 11592 (2021).<https://doi.org/10.1038/s41598-021-90063-3>
- Zhang GP (2003). Time series forecasting using a hybrid ARIMA and neural network model.*Neurocomputing*, 20 159-175.

Appendix

A.1

R Code

```
library(readxl)
read_excel("C:\\Users\\ADMIN\\Desktop\\paper1\\Paper1_phd\\data_survey1_modified.xlsx")
summary(data_survey1_modified)
library(quantreg)
t<- matrix(table(data_survey1_modified$Impact_score),ncol=8)
q<-t[,1]/sum(t);q
q_1<- the value of median quantile is taken from [13]
q_2<- median quantile
f_0e<-( q_1*q)/q_2; f_0e
##Null Model
model.qr_null<-rq(formula = Impact_score~1,tau=q,data = data_survey1_modified)
summary(model.qr_null)
s1<-sum(abs(model.qr_null$residuals));s1
#####Full model
model.qr3<-rq(formula=Impact_score~HyperactivityScale+Emotional+Conduct_Problem+Peer_Problem,tau=q,data =
data_survey1_modified)
summary(model.qr3)
s2<-sum(abs(model.qr3$residuals));s2
R_q1 <- 1-(s2/s1); R_q1
##H_0<- There is no significant difference between null and complete model
##H_1<- There is significant difference between null and complete model
LR_1<-2*(s1-s2)*f_0e/(q*(1-q));LR_1
```

```
summary(model.qr_null,se="boot")
summary(model.qr3,se="boot")
fit_qr3<-model.qr3$fitted.values
Final_fit<-ifelse(fit_qr3<0.5,"Normal",ifelse(fit_qr3>=1.5,"Abnormal","Borderline"))
```

A.2 Quantile Regression Neural Network (QRNN) Model

```
###Firstly the whole data divided into Train and Test data (60% Train and 40% Test)
Train_b1<-sample(60% of survey first data,replace = FALSE)
Train_quant1<-data_survey1_modified[Train_b1,]
Test_quant1<-data_survey1_modified[-Train_b1,]
###QR model
library(quantreg)
model.qr_train1<-rq(formula=Impact_score~HyperactivityScale+Emotional+Conduct_Problem+Peer_Problem ,tau=q,data
=Train_quant1)
fit_qr_train1<- model.qr_train1$fitted.values
Final_fit_train1<-ifelse(fit_qr_train1<0.5,"Normal",ifelse(fit_qr_train1>=1.5,"Abnormal","Borderline"))
summary(model.qr_train1,se="boot")
library(qrnn)
diff_compo1<-
as.matrix(data.frame(Train_quant1$HyperactivityScale,Train_quant1$Emotional,Train_quant1$Conduct_Problem,Train_qua
nt1$Peer_Problem),ncol=4)
test_compo1<-
as.matrix(data.frame(Test_quant1$HyperactivityScale,Test_quant1$Emotional,Test_quant1$Conduct_Problem,Test_quant1$
Peer_Problem),ncol=4)
ml_qrnn2<-qrnn.fit(diff_compo1,as.matrix(Train_quant1$Impact_score),tau = q,n.hidden = 1)
t_tau1<-seq(0.40,0.48,0.02)
x.y.tau1 <- composite.stack(diff_compo1,as.matrix(Train_quant1$Impact_score) ,t_tau1)
binary.tau1 <- dummy.code(as.factor(x.y.tau1$tau))
fit.cqr1 <-qrnn.fit(cbind(binary.tau1, x.y.tau1$x), x.y.tau1$y,tau=x.y.tau1$tau, n.hidden=1, n.trials=1,Th=linear,
Th.prime=linear.prime)
pred.cqr1 <- matrix(qrnn.predict(cbind(binary.tau1, x.y.tau1$x), fit.cqr1),ncol=length(t_tau1))
Final_fit_qrnn<-ifelse(pred.cqr1<0.5,"Normal",ifelse(pred.cqr1>=1.5,"Abnormal","Borderline"))
coef.cqr<- lm.fit(cbind(1,diff_compo1), pred.cqr1)$coef; coef.cqr
### train data the value of q is 0.43, hence same value applied for test data
diff_compo2<-
as.matrix(data.frame(Train_quant1$HyperactivityScale,Train_quant1$Emotional,Train_quant1$Conduct_Problem,Train_qua
nt1$Peer_Problem),ncol=4)
test_compo2<-
as.matrix(data.frame(Test_quant1$HyperactivityScale,Test_quant1$Emotional,Test_quant1$Conduct_Problem,Test_quant1$
Peer_Problem),ncol=4)
t1<- matrix(table(Train_quant1$Impact_score),ncol=7)
q_1<-t1[,1]/sum(t1);q_1
t_tau2<-q_1
ml_qrnn1<-qrnn.fit(diff_compo2,as.matrix(Train_quant1$Impact_score),tau = q_1,n.hidden = 1)
x.y.tau2 <- composite.stack(diff_compo2,as.matrix(Train_quant1$Impact_score) ,t_tau2)
binary.tau2 <- dummy.code(as.factor(x.y.tau2$tau))
fit.cqr2 <- qrnn.fit(cbind(binary.tau2, x.y.tau2$x), x.y.tau2$y,
tau=x.y.tau2$tau, n.hidden=1, n.trials=1,
Th=linear, Th.prime=linear.prime)
pred.cqr2 <- round(matrix(qrnn.predict(cbind(binary.tau2, x.y.tau2$x), fit.cqr2),
ncol=length(t_tau2)),3)
pred.cqr2 <- matrix(qrnn.predict(cbind(binary.tau2, x.y.tau2$x), fit.cqr2),ncol=length(t_tau2))
Final_fit_qrnn_0.43<-ifelse(pred.cqr2<0.5,"Normal",ifelse(pred.cqr2>=1.5,"Abnormal","Borderline"))

####test data q=0.43
test_compo3<-
as.matrix(data.frame(Test_quant1$HyperactivityScale,Test_quant1$Emotional,Test_quant1$Conduct_Problem,Test_quant1$
Peer_Problem),ncol=4)
```

```
ml_qrnn1<-qrnn.fit(test_comp3,as.matrix(Test_quant1$Impact_score),tau =q_1,n.hidden = 1)
t_tautest1<-q_1
x.y.tautest1 <- composite.stack(test_comp3,as.matrix(Test_quant1$Impact_score) ,t_tautest1)
binary.tautest1 <- dummy.code(as.factor(x.y.tautest1$tau))
fit.cqrtest1 <- qrnn.fit(cbind(binary.tautest1, x.y.tautest1$x), x.y.tautest1$y,
tau=x.y.tautest1$tau, n.hidden=1, n.trials=1,
Th=linear, Th.prime=linear.prime)
pred.cqrtest1 <- round(matrix(qrnn.predict(cbind(binary.tautest1, x.y.tautest1$x), fit.cqrtest1),
ncol=length(t_tautest1)),3)
pred.cqrtest1<-matrix(qrnn.predict(cbind(binary.tautest1,x.y.tautest1$x),fit.cqrtest1),ncol=length(t_tautest1))
Final_fit_qrnn_test<-ifelse(pred.cqrtest1<0.5,"Normal",ifelse(pred.cqrtest1>=1.5,"Abnormal","Borderline"))

###Comparison of Both Models by MAE, MSE and RMSE for different values of tau
### MAE, MSE and RMSE for QR Model
MAE_qr<-mean(abs(Train_quant1$Impact_score - model.qr_train1$fitted.values ))
MSE_qr<-mean((model.qr_train1$fitted.values - Train_quant1$Impact_score)^2)
RMSE_qr<-sqrt(mean((model.qr_train1$fitted.values - Train_quant1$Impact_score)^2))
### MAE, MSE and RMSE for QRNN Model
MAE_qrnn<-mean(abs(Train_quant1$Impact_score - pred.cqr1[,3] )) ##For tau=0.44
MSE_qrnn<-mean((pred.cqr1[,3] - Train_quant1$Impact_score )^2)
RMSE_qrnn<-sqrt(mean((pred.cqr1[,3] - Train_quant1$Impact_score )^2))
```
