IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/rjes

Semi-Parametric Fuzzy Quantile Regression Model Estimation Based on Proposed Metric via Jensen–Shannon Distance

Elaf Baha Alwan and Omar Abdulmohsin Ali

Department of Statistics College of Administration and Economics, University of Baghdad, 10051 Baghdad, Iraq.

ARTICLE INFO

Article history:
Received 13/11/2024
Revised 13/11/2024
Accepted 15/01/2025
Available online 15/05/2025

Keywords:
Fuzzy Metric,
Shannon Entropy,
Quantile Regression,
Semiparametric Estimation,
Artificial Intelligence

ABSTRACT

Fuzzy regression is considered one of the most important regression models, and recently the fuzzy regression model has become a powerful tool for conducting statistical operations, however, the above model also faces some problems and violations, including (when the data is skewed, or no-normal, ...) and thus leads to incorrect results, so it is necessary to find a model to deal with such violations and problems suffered by the regular fuzzy regression models and at the same time be more powerful and immune than the fuzzy regression model called the semi-parametric fuzzy quantile regression. This model is characterized by containing two parts, the first is the fuzzy parametric part (fuzzy inputs and crisp parameters) and the second is the fuzzy nonparametric part for fuzzy triangular numbers, and the semiparametric fuzzy quantile regression is estimated. To demonstrate the effectiveness of our combining model, we will utilize the following Akbari and Hesamian (2019) dataset that was used as a reference case study. Estimate Fuzzy Quantile Regression Model: (FQRM), Fuzzy semi-parametric quantile regression: (FSPQRM), Fuzzy Support Vector Machine: (FSVM), Combining FQRM-FSVR (Comb), Combining FSPQRM-FSVR. Using a new metric measure Jensen–Shannon Distance: (JS) based on fuzzy belonging functions. Two criteria MSM and G1 were used in comparison.

1. Introduction

The first goal of using fuzzy regression models is the need to handle uncertainty in both the dependent and independent variables, especially when data cannot be expressed with precise numerical values. This is where fuzzy regression comes into play—allowing for the incorporation of fuzzy numbers (which represent ranges of values rather than specific points) into the regression framework. However, despite advances in fuzzy regression methods, there are still significant limitations in handling quantile regression in the context of fuzzy data Quantile regression has become a powerful tool in statistical modeling, especially when dealing with non-normal data distributions or when focusing on specific quantiles (e.g., median, upper, or lower quantiles) Koenker, R. (2005). rather than the mean. This approach is particularly useful in contexts where the relationship between variables is non-linear or asymmetric, which is common in many real-world datasets. However, traditional regression models, especially those relying on crisp or deterministic data, often fail to capture the inherent uncertainty and vagueness present in real-world data. This is especially true for datasets with fuzzy or imprecise information, where values are not exactly known but rather expressed as fuzzy numbers or intervals

This paper aims to propose a semi-parametric fuzzy quantile regression model that estimates the relationships between fuzzy variables in a way that captures both the non-linearity and uncertainty inherent in real-world datasets. The semi-parametric approach is used to combine both parametric (linear) and non-parametric

Corresponding author: Leccit2@uowasit.edu.iq
https://doi.org/10.62933/3e7qvw48
This work is licensed under https://creativecommons.org/licenses/by-nc-sa/4.0/

(flexible, data-driven) components in the model. This method allows for the estimation of quantiles of the fuzzy output variable, providing a more robust framework for modeling the relationship between fuzzy predictors and the response.

A central innovation in this study is the use of the Jensen–Shannon (JS) distance as a metric for comparing fuzzy distributions. The JS distance is a symmetric measure of the similarity between two probability distributions and is particularly useful for measuring the difference between fuzzy distributions, making it a perfect fit for fuzzy quantile regression models. This metric is introduced to quantify the distance between fuzzy numbers representing the actual data and the predicted outcomes in the regression model.

Jensen–Shannon Distance allows the model to handle uncertainty more effectively, as it accounts for the spread or vagueness in fuzzy data. The proposed metric ensures that the quantiles of the fuzzy response are estimated with greater precision, even in the presence of noise or outliers.

2. Methodology

2.1 Shannon entropy

Fuzzy entropy refers to the quantity of uncertain information that can be obtained from a fuzzy set or fuzzy system. Specifically, it should be noted that fuzzy entropy is defined without the necessity for a probabilistic notion, which sets it apart from the traditional Shannon entropy. The rationale behind this is that fuzzy entropy incorporates uncertainties related to vagueness and ambiguity, whereas Shannon entropy only includes probabilistic randomness uncertainties (Al-sharhan 2001; Arora 2021). The membership function is used to define the fuzzy entropy. (De Luca and Termini,1993) presented a set of conditions that a fuzzy entropy should satisfy and define fuzzy entropy based on Shannon's function in 1972. The characteristics of fuzzy entropy are commonly acknowledged and have evolved into a standard by which new fuzzy entropies are defined.

Equation (1) displays the fuzzy entropy that (DeLuca and Termini, 1993) proposed its definition is predicated on the idea of a membership function, of which there are $\mu\,\tilde{N}\,(u_i$. It is widely used in various fields, including communication theory, data compression, cryptography, and statistical physics. Let $\tilde{N}$ is LR-FN the Shannon entropy of $\tilde{N}$ are define as:

$$ent(\tilde{N}) = \sum_{i=1}^{n} S(\mu\,\tilde{N}\,(u_i)) \quad (1)$$

where:

$$S\left(\mu\,\tilde{N}\,(u_i)\right) = -u\log(u) - (1 - u)\log(1 - u) \quad (2)$$

$$KL(N, M) = \sum_{i=1}^{n} ai \log(\frac{ni}{mi}) \quad (3)$$

$$KL(\tilde{N}, \tilde{M}) = \int \mu\tilde{N} \log(\frac{\mu\,\tilde{N}}{\mu\,\tilde{M}}) \quad (4)$$

Where we consider that

$0 \cdot \ln(0) = 0$ and $\ln(0)/0 = 0$.

2.2 Jensen–Shannon Distance

Let

$\tilde{N} = \{\langle u, \mu\,\tilde{N}\,(u)\rangle | u \in U\}$ and $\tilde{M} = \{\langle u, \mu\,\tilde{M}\,(u)\rangle | u \in U\}$ are fuzzy numbers, Jensen–Shannon distance $DJS(\mu_n, \mu_m)$ is defined as:

$$DJS(\mu_n, \mu_m) = \frac{1}{n}[KL1(\mu_n, v) + KL2(\mu_m, v*) \quad (5)$$

where:

$$KL1(\mu_N, v) = \frac{1}{2}\sum_{i=1}^{n} \mu_N(u_i)v \quad (6)$$

$$KL2(\mu_M, v*) = \frac{1}{2}\sum_{i=1}^{n} \mu_M(u_i)v* \quad (7)$$

$$v = log\frac{\mu_N(u_i)}{\mu_M(u_i)} \quad (8)$$

$$v* = log\frac{(1-\mu_N(u_i))}{(1-\mu_M(u_i))} \quad (9)$$

where: KL is Kullback-Leibler divergence, and JS is Jensen-Shannon.

Properties:

• DJS($\widetilde{N},\widetilde{M}$) =0 if $\widetilde{M}=\widetilde{N}$

• $0 \le$ DJS($\widetilde{N},\widetilde{M}$) ≤ 1

• D($\widetilde{N},\widetilde{M}$) $\le$ DJS($\widetilde{N},\widetilde{Z}$) + DJS($\widetilde{Z},\widetilde{M}$) if $\widetilde{N} \subseteq \widetilde{M} \subseteq \widetilde{Z}$

2.3 FSPQRM Based on Jensen–Shannon Divergence Distance

Jensen-Shannon Divergence in (2.2), is utilized for the FSPQRM model estimation in where, the optimization problem associated with FSPQRM based on Jensen–Shannon Divergence is formulated as equation (10).

$$\min_{\beta\in R^k} V\big(\beta(\tau)\big) = \min_{\beta\in R^k}[\frac{1}{n}\sum_{i=1}^{n} \rho\tau \ \mathrm{DJS}\big(\mu_{\tilde{y}_i}, \mu_{\hat{\tilde{y}}_i}\big) \qquad (10)$$

appears to be the most optimal choice among the tested values. Therefore, based on the provided results, Can conclude that τ = (0.2 , 0.8) is the best choice for achieving the highest performance in terms of both G1 and MSM metrics. From the provided table, you see that the number of iterations varies for each τ value. For instance, the number of iterations ranges from 14 to 18 across the different τ values tested. This variation in the number of iterations suggests differences in the convergence behavior of the optimization algorithm for different τ values. The number of iterations required for convergence can be an important consideration in practical optimization tasks. In general, a higher number of iterations may indicate that the optimization process is taking longer to converge to the optimal solution. Conversely, a lower number of iterations may suggest faster convergence. In the context of this specific optimization problem, it's essential to strike a balance between the number of iterations and the achieved performance metrics (such as G1 and MSM). While a higher number of iterations may lead to potentially better convergence and finer optimization of the objective function, it

So our proposed model is derfrint of previous one that introduced by (Akbari and Hesamian 2019; Hesamian and et al 2017), where they work are based on the α –value metric, weare based on membership function, our proposed of this metric are get model with less complexity.

3. Results and discussion

3.1. FQRM results

Through table (1, 2 , 3: FQRM on JS metric),each row represents the results obtained for a specific value of the parameter τ By examining the values in the G1 and MSM, can observe that the highest values are obtained for τ = 0.2 At this τ value, the G1 metric reaches its maximum value of approximately (0.725372919) when N=10 , and the MSM metric also reaches its maximum value of around (0.704490731) when N=10 and τ =0.8 . is suggests that for this particular dataset or problem, τ = (0.2 , 0.8).

also increases computational costs. On the other hand, a lower number of iterations may result in faster computations but could potentially compromise the accuracy of the optimization results. Therefore, when considering the number of iterations, it's crucial to evaluate it alongside the achieved performance metrics and computational resources available. In practical applications, one might choose a τ value that not only maximizes performance metrics like (G1 and MSM) but also converges within a reasonable number of iterations, balancing computational efficiency with optimization accuracy.

The proposed distance function (JS) was also applied to the model (FQRM) in table (1) using sample size (N=10) and (τ =0.2, 0.4, 0.6, 0.8) and according to the criteria MSM , G1 used mentioned in chapter two, the best result was as follows: (G1=0.72537), when τ =0.2 MSM=0.704490731) when τ =0.8 .

Table 1: FQRM for different values of (τ) based on JS metric

Performance	When N=10			
	τ =0.2	τ =0.4	τ =0.6	τ =0.8
MSM	0.70445	0.70432	0.70413	0.70449
G1	0.72537	0.72125	0.72370	0.72492
Beta/ hat	0.11557 1.66707	0.12669 1.67004	0.05259 1.70361	1.00000 1.7069
Itr	14	12	14	18

Table 2: FQRM for different values of (τ) based on JS metric

Performance	When N=20			
	τ =0.2	τ =0.4	τ =0.6	τ =0.8
MSM	0.43038	0.46215	0.46217	0.46221
G1	0.49984	0.46871	0.46881	0.46881
Beta/ hat	0.10483 1.49211	0.99325 1.10494	0.98941 1.10559	0.98422 1.10552
Itr	**16**	**18**	**20**	**20**

The proposed distance function (JS) was also applied to the model (FQRM) in table (2) using sample size (N=20) and (τ =0.2, 0.4, 0.6. 0.8) and according to the criteria used MSM, G1 mentioned in chapter two, the best result was as follows: (G1=0.499838497) when τ =0.2, (MSM=0.462213225) when τ =0.8.

Table 3: FQRM for different values of (τ) based on JS metric

Performance	When N=50			
	τ =0.2	τ =0.4	τ =0.6	τ =0.8
MSM	0.48298	0.49008	0.48979	0.490659
G1	0.53655	0.50985	0.50889	0.513052
Beta/ hat	0.10216 1.49370	0.63969 1.25708	0.68357 1.25112	0.65091 1.27449
Itr	18	20	18	22

The proposed distance function (JS) was also applied to the model (FQRM) in table (3) using sample size (N=50) and (τ =0.2, 0.4, 0.6. 0.8) and according to the criteria used MSM , G1 mentioned in chapter two, the best result was as follows: (G1=0.536546757) when τ =0.2 , (MSM=0.49065963) when τ =0.8 .

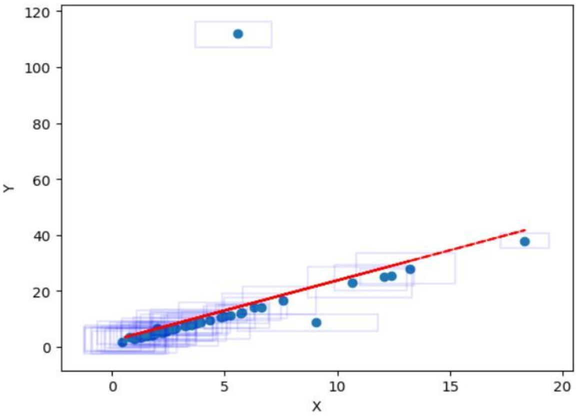

Figure 1: Akbari Data Visualization

3.2 Nonparametric FSVR results

When running the fuzzy support vector machine regression (FSVR) model with the provided equation: Hong & Hwang (2003), Smola & Schölkopf (2004).

$$\tilde{y}_{FSVR} = 2.075968787386472\,\tilde{x} + (1.07296658;\ 2.08782432, 2.66640378) \quad (11)$$

We obtained performance metrics G1 and MSM with the following values table (4). These metrics are crucial for evaluating the effectiveness of the FSVR model in capturing the relationship between the input variable x̃ and the output variable Y . The G1 metric measures the goodness-of-fit of the model to the observed data. A value closer to indicates a better fit, suggesting that the FSVR model explains a significant portion of the variance in the data. In this case, tabel (3.4) when n=10 the obtained G1 value of approximately 0.773827 suggests a reasonably good fit of the FSVR model to the data. The MSM metric evaluates the mean squared error between the observed and predicted values. A lower MSM value indicates better predictive performance of the model. The obtained MSM value of around 0.748672 suggests that the FSVR model's predictions are relatively close to the observed values on average. So, based on the provided performance metrics, the FSVR model seems to provide a reason- ably good fit to the data and demonstrates adequate predictive accuracy. However, further analysis and comparison with alternative models may be necessary to fully assess Their effectiveness and suitability for the given task or application.

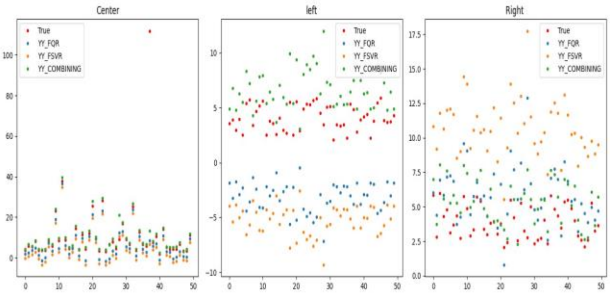

Figure 3: Akbari Data Visualization with Prediction with Scatter

3 FSPQRM results

Table (3.3) shows the outcomes of the FSPQRM approach, focusing on signed Dk, across various τ values and kernel types. The table is organized according to the metric utilized, where ”JS” denotes Jensen- Shannon divergence, and ”AK” represents the Dk that is defined $D_k(\widetilde{N},\widetilde{M}) = \int \quad w_k(\alpha)\,[\widetilde{N}_\alpha - \widetilde{M}_\alpha)d\alpha$

Upon examining the data, several noteworthy observations emerge regarding the estimated parameters (βˆ(τ)) and performance metrics (MSM and G1) across different τ values and kernel types:

ˆ Under the Jensen-Shannon divergence (JS) metric:

–The estimated parameters $\hat{\beta}_{(\tau)}$ exhibit variability within the range of 1.993 to 2.039, indicating diverse levels of captured information by the model.

–The mean squared error (MSM) spans approximately from 0.5019 to 0.5952, reflecting the model’s accuracy in fitting the data.

–Similarly, the G1 statistic fluctuates between 0.6451 and 0.7042, signifying the goodness-of-fit of the model.

ˆ Under the (AK) metric:

–The estimated parameters $\hat{\beta}_{(\tau)}$ also demonstrate variability within a comparable range as observed in the JS metric.

–The MSM values range from roughly 0.5009 to 0.5948, with similar interpretations to those observed in the JS metric.

–The G1 statistic spans from 0.6442 to 0.7048, indicating the model’s fit to the data.

Overall, these results underscore the efficacy of the FSPQRM method in capturing the unlearned portion of the data distribution across different τ values and kernel types, as evidenced by the variations in estimated parameters and performance metrics.

Furthermore, the number of iterations (itr) is a noteworthy aspect to discuss. Generally, higher numbers of iterations may imply that the optimization algorithm required more steps to converge to a solution, while lower numbers suggest faster convergence. The choice of metric can influence the number of iterations, as observed with JS typically requiring fewer iterations compared to AK. This discrepancy could stem from various factors such as the optimization landscape, algorithm suitability, and convergence criteria. It’s crucial to strike a balance between computational efficiency and optimization performance when selecting metrics for the FSPQRM model.

Table 4: FSVMr (Fuzzy Support Vector Machines Regresson)

Preformance	N=10	N=20	N=50
W	0.59733	0.49457	2.0.532
B	1.10656	1.02604	1.31644
L_B	2.13141	2.49044	1.82277
R_B	1.56915	2.59679	2.01128
MSM	0.74867	0.76271	0.77474
G1	0.773828	0.78042	0.793525

Through table (4) we notice that the (G1) criterion leads to better results than the (MSM) criterion when using different sample sizes (N=10, 20, 50) and depending on the weights as well as (W) (R_B, L_B), where the value of (G1=0.7738275111), (MSM=0.748672438) when N=10 and (G_1=0.7804235134), (MSM=0.762711). when N=20 and it also leads to better and faster results than the previous methods (FQRM) can be relied upon.

3.4 COMBINING Models results

In this section two COMBINING Models: COMBINING FQRM-FSVR and COMBINING FSPQRM- FSVR, as follows: [Burman & Chaudhuri (2012)].

3.4.1 COMBINING FQRM-FSVR Model

When combining the FQRM-FSVR models using the equation:

$$\hat{y}_{\text{COMBINING}} = (1 - \pi)\hat{y}_{FQRM} + \pi\hat{y}_{FSVR} \qquad (12)$$

where π= represents a weighting parameter, we obtained performance metrics G1 and MSM with the following values:

G1=0.7 (13)

MSM= 0.57 (14)

3.4.2 COMBINING FSPQRM-FSVR Model

Like previous case when combining the FSPQRM-FSVR models using the equation:

$$\hat{y}_{\text{COMBINING}} = (1 - \pi)\hat{y}_{FSQRM} + \pi\hat{y}_{FSVR} \quad (15)$$

where π= represents a weighting parameter [Stein (1956, January)], we obtained performance metrics G1 and MSM with the following table (5):

Table 5 COMBINING FSPQRM-FSVR Model

Performance	G_1	MSM
N=10	0.848126743	0.83946839
N=20	0.827928938	0.795617246
N=50	0.841730651	0.826087104

Table 5 COMBINING FSPQRM-FSVR Model When B=1

Performance	G_1	MSM
N=10	0.848126743	0.83946839
N=20	0.827928938	0.795617246
N=50	0.841730651	0.826087104

4. Conclusions

The researcher reached the following main points:

- When comparing the distance functions, he concluded that the proposed function (JS) is better using the mentioned comparison criteria, and the number of iterations is less, and thus the convergence speed is faster using different sample sizes(10,20,50).

References

[1] Al-Sharhan, S., Karray, F., Gueaieb, W., & Basir, O. (2001, December). Fuzzy entropy: a brief survey. In *10th IEEE international conference on fuzzy systems.(Cat. No. 01CH37297)* (Vol. 3, pp. 1135-1139). IEEE.

[2] Arora, M., and Kumar, R. (2021). Fuzzy c-means clustering algorithm with entropy-based initialization for medical image segmentation. Soft Computing, 25(10), 7869-7882.

[3] De Luca, A., & Termini, S. (1993). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. In *Readings in fuzzy sets for intelligent systems* (pp. 197-202). Morgan Kaufmann.

[4] Hesamian, G., & Akbari, M. G. (2017). Semi-parametric partially logistic regression model with exact inputs and intuitionistic fuzzy outputs. *Applied Soft Computing*, *58*, 517-526.

[5] Hesamian, G., & Akbari, M. G. (2019). Fuzzy quantile linear regression model adopted with a semi-parametric technique based on fuzzy predictors and fuzzy responses. *Expert systems with applications*, *118*, 585-597.

[6] Hong, D. H., & Hwang, C. (2003). Support vector fuzzy regression machines. *Fuzzy sets and systems*, *138*(2), 271-281.

[7] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*, 199-222.

[8] Stein, C. (1956, January). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 1, pp. 197-206).

[9] Burman, P., & Chaudhuri, P. (2012). On a hybrid approach to parametric and nonparametric regression. In *Nonparametric Statistical Methods and Related Topics: A Festschrift in Honor of Professor PK Bhattacharya on the Occasion of his 80th Birthday* (pp. 233-256).

[10] Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge university press.