Vol 5, Issue 1, January 2025 www.ijesti.com E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

Comparative Analysis of Regression Models for Quantile Prediction Using Pinball Loss

V Rajanikanth Tatiraju [1]

[1] Research Scholar, Department of Computer Science and Engineering,
P. K. University, Shivpuri, M.P.

Dr. Rohita Yamaganti [2]

[2] Associate Professor, Department of Computer Science and Engineering,
P. K. University, Shivpuri, M.P.

ABSTRACT

The performance of three regression models, namely Lagrangian Asymmetric-vTwin Support Vector Regression (SVR), Standard SVR, and Linear Regression, is examined and compared in this study. The models are tested using various quantiles of Pinball Loss, $\alpha = 0.1$, 0.5, and 0.9, in addition to more conventional metrics such as RMSE and MAE. Pinball Loss values specific to each quantile were used to evaluate the models' performance after training and testing on a regression dataset to forecast the lower, median, and higher quantiles. The outcomes show that Lagrangian Asymmetric-vTwin SVR is the best option, providing the lowest Pinball Loss, RMSE, and MAE, compared to Standard SVR and Linear Regression. Additionally, it was discovered that the ideal C value, which is 1.0, successfully balanced training duration and prediction accuracy.

Key Words: *Pinball, Quantile, Lagrangian, Asymmetric, Performance.*

I. INTRODUCTION

The establishment of links between dependent and independent variables is a crucial step in predictive modeling, and regression analysis plays a key role in this process. Predicting the dependent variable's mean from the independent variables is the main emphasis of most regression models in the past. This method is called conditional mean estimation. In many real-world situations, though, this assumption might not be enough; for example, if the data shows strong tails or skewness, or if you need more specifics regarding the distribution of the target variable for your decision-making. In response to these issues, quantile regression has developed into a strong substitute that enables the prediction of different quantiles of the response variable's conditional distribution. By estimating the mean and other features of the distribution, such the behavior of the tails, this gives a more complete picture of the data.

The method of estimating the conditional quantiles of a response variable in relation to predictor factors is known as quantile regression. It was initially proposed by Koenker and Bassett in 1978. The goal of quantile regression is to minimize a weighted sum of absolute residuals, where the weights are determined by the quantile of interest, as opposed to ordinary least squares (OLS), which minimizes the sum of squared residuals. For data with an asymmetric distribution or predictors with varying impacts across quantiles, this method shines. In economic data, for instance, quantile regression is useful for evaluating the heterogeneous impacts of variables since the link between income and education may differ for low-income and high-income individuals.

In contexts where different quantiles (e.g., the 90th or 10th percentile) may hold different significance, such as risk management, medical studies, and climate modeling, quantile regression's capacity to offer a more comprehensive characterization of the dependent variable's conditional distribution becomes extremely important. The asymmetry of the quantiles must be taken into consideration by the loss function for quantile regression to be effective, though. Here we see the application of pinball loss, a loss function that quantile regression models have come to embrace.

In quantile regression, the pinball loss—sometimes called the tilted absolute loss—is the best fit because it penalizes overestimations and underestimations in an asymmetrical fashion. In particular, it enables the model to highlight prediction mistakes in a different way whether the quantile is higher or lower than the actual number. Pinball loss's computational speed and flexible handling of various data types, including those with non-normal distributions or heterogeneity in the errors, have led to its increasing adoption. Quantile regression is now more approachable for issues of a large scale as the loss function is a part of many machine learning techniques. It is compatible with regularization methods like L1 and L2, which help prevent overfitting and guarantee robust model predictions, and it may be used in conjunction with optimization approaches like gradient descent.

Pinball loss is popular in regression models that rely on deep learning in part because of how versatile it is. Combining neural networks with pinball loss allows for quantile regression in many different contexts, because to neural networks' ability to capture complicated correlations between variables. Fields like healthcare, where forecasting the upper quantile of a variable like patient recovery time can have critical implications for resource allocation, and finance, where models need to predict the tail risks (such as extreme market crashes or booms) have both benefited greatly from this approach.

In time series forecasting, quantile regression with pinball loss has been used to predict quantiles of future values, which is useful because the data is frequently non-stationary and autocorrelated. Important for making decisions when faced with uncertainty, this enables the modeling of prediction uncertainty. In energy consumption forecasting, for example, it may be more useful to anticipate the 95th percentile of future demand than the mean, as this information is useful for choices about infrastructure capacity and load balancing.

When dealing with heteroscedasticity, a key component of quantile regression models based on pinball loss is ensuring that the variance of the error components remains consistent across data. If this is the case, it's possible that the uncertainty in the predictions won't be reliably estimated using conventional regression methods like ordinary least squares (OLS). Pinball loss regression, on the other hand, can improve the model's performance in cases of uneven variability by concentrating on quantiles, giving more robust and accurate predictions throughout the distribution.

II. REVIEW OF LITERATURE

Sigauke, Caston et al., (2018) Using additive quantile regression (AQR) models, this paper discusses short-term hourly load forecasting in South Africa. By using this method, the combined modeling of hourly power data is easily interpretable and takes residual autocorrelation into consideration. The use of generalised additive models (GAMs) allows for a comparative examination. Hierarchical interactions are used in both modeling frameworks to choose variables using the least absolute shrinkage and selection operator (Lasso). Each of the four models—GAMs with interactions and AQR models without—are carefully examined. The most accurate model that suited the data best was the AQR model that included pairwise interactions. Quantile regression averaging (QRA) and an algorithm based on the pinball loss

(convex combination model) were used to integrate the forecasts from the four models. After comparing the AQR model with interactions to the convex combination and QRA models, it was found that the QRA model produced the best accurate forecasts. Both the convex combination model and the QRA model, with the exception of the AQR model with interactions, provided appropriate prediction interval coverage probabilities for the 90%, 95%, and 99% intervals. In terms of average width and average deviation normalized by prediction interval, the QRA model was the most compact. Going beyond summary performance statistics in forecasting has benefit, as it offers additional insight into the built forecasting models. This can be seen in the modeling framework mentioned in this study.

Yu, Lean et al., (2018) The development of new quantile estimators and a loss function that takes into account the noise in both the response and explanatory variables allows for reliable quantile estimations to be achieved, even in the presence of noisy data. This is especially true when orthogonal loss is substituted for vertical loss in conventional quantile estimators, resulting in an improvement over pinball loss called orthogonal pinball loss (OPL). In this way, new OPL-based QR and SVMQR models may be developed from existing linear and support vector machine quantile regression programs, respectively. In terms of quantile property and prediction accuracy, the empirical analysis on 10 publicly accessible datasets statistically confirms that the two OPL-based models outperform their respective original forms, particularly for extreme quantiles. An innovative OPL-based SVMQR model that incorporates AI achieves better results than any benchmark model; this makes it a potentially useful quantile estimator, particularly when dealing with noisy data.

Hu, Ting et al., (2012) A kernel-based online learning technique linked to a series of insensitive pinball loss functions is being considered for use in quantile regression and support vector regression. The quantile parameter $\tau\tau$ has the potential to affect the statistical performance of the learning algorithm, as demonstrated quantitatively by our error analysis and derived learning rates. We successfully navigated the technical challenge posed by the sparsity-motivated introduction of a variable insensitive parameter in our analysis.

Steinwart, Ingo & Christmann, Andreas. (2011) A popular method in machine learning and statistics, the so-called pinball loss estimates conditional quantiles. The effectiveness of this tool for nonparametric techniques, however, has received surprisingly little attention thus far. To address this, we prove certain inequality that characterize the proximity of the approximate pinball risk minimizers to the relevant conditional quantile. These disparities, which persist under modest assumptions on the distribution of the data, are then utilized to construct so-called variance limits, which have lately emerged as crucial tools in the statistical evaluation of (regularized) empirical methods for minimizing risk. Lastly, we prove an oracle inequality for SVMs using the pinball loss by combining the two kinds of inequalities. With respect to the conditional quantile, the ensuing learning rates are min-max optimum under certain conventional regularity assumptions.

Zheng, Songfeng. (2011) It is common for optimization algorithms that rely on gradients to rapidly converge to a local maximum. Unfortunately, the quantile regression model's use of a check loss function that isn't always differentiable rules out the use of gradient based optimization techniques. Therefore, in order to fit the quantile regression model using gradient based optimization methods, this study presents a smooth function to approximate the check loss function. We go over the features of the smooth approximation. The objective function that has been smoothed can be minimized using two different approaches. Two methods have been developed for smooth quantile regression: one uses gradient descent directly, which produces the gradient descent smooth quantile regression model; the other uses functional

gradient descent to minimize the smoothed objective function; and finally, boosted smooth quantile regression algorithm is the result of changing the fitted model along the negative gradient direction in each iteration. The suggested smooth quantile regression algorithms outperform other quantile regression models in terms of prediction accuracy and efficiency in eliminating noninformative variables, according to extensive tests conducted on both real-world and simulated data.

Somers, Mark & Whittaker, Joe. (2007) Two examples of retail credit risk assessment using quantile regression show how the method can handle the wide range of distributions seen in the banking sector. One use case is in the prediction of loss due to default for secured loans, namely residential mortgages. Banks do not keep the profit when the value of the security (such a property) exceeds the loan balance; conversely, they incur a loss when the value of the security falls short of the defaulting debt. This creates an asymmetric process. Because of this imbalance, it's clear that evaluating the house's low end value—where losses are most likely to occur—is far more useful for this purpose than calculating the average value, which seldom experiences losses. In our application, we estimate the distribution of property values realized upon repossession using quantile regression. This distribution is then utilized to quantify loss given default estimations. A mortgage lender in Europe provides an example of their portfolio. Another area where it finds use is in revenue modeling. Credit granting organizations have access to massive information, but they also create models to predict how new tactics will play out, even while there is inherently no evidence available for such techniques. In certain markets, the goal of implementing a strategy is to either increase revenue or decrease risk. To better understand which accounts are most and least lucrative based on their anticipated variables, we use quantile regression in a basic artificial revenue model. Kernel smoothed quantile regression and conventional linear regression are employed in the application.

III. EXPERIMENTAL SETUP

In this study, the performance of three regression models—Lagrangian Asymmetric-vTwin Support Vector Regression (SVR), Standard SVR, and Linear Regression—will be evaluated and compared. This will be done using different quantiles of Pinball Loss ($\alpha = 0.1$, 0.5, and 0.9), as well as other metrics such as RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). Every model is trained and tested on a regression dataset, and its performance is evaluated based on how well it can predict lower, median, and higher quantiles (Pinball Loss values for $\alpha = 0.1$, 0.5, and 0.9). Furthermore, the SVR models are fine-tuned by adjusting the regularization parameter, C, to the following values: 0.1, 1.0, 10.0, and 100.0. The impact of these adjustments on RMSE, MAE, and Pinball Loss (when $\alpha = 0.5$) is examined, as well as the amount of time it takes to train each configuration.

IV. RESULTS AND DISCUSSION

Table 1: Model Performance with Different Pinball Loss Quantiles

Model	Pinball Loss ($\alpha = 0.1$)	Pinball Loss ($\alpha = 0.5$)	Pinball Loss ($\alpha = 0.9$)	RMSE	MAE
Lagrangian Asymmetric-vTwin SVR	0.070	0.082	0.095	0.252	0.181
Standard SVR	0.090	0.105	0.112	0.297	0.210
Linear Regression	0.110	0.120	0.132	0.335	0.233

The table shows the performance of three models—Lagrangian Asymmetric-vTwin SVR, Standard SVR, and Linear Regression—using different pinball loss quantiles (α = 0.1, 0.5, 0.9), as well as RMSE and MAE. The Lagrangian Asymmetric-vTwin SVR model consistently outperforms the other models. It has the lowest overall loss values and error metrics across all quantiles (0.070, 0.082, 0.095 for α = 0.1, 0.5, and 0.9, respectively) and has the lowest RMSE (0.252) and MAE (0.181). Standard SVR performs better than Linear Regression, however it still does not perform as well as the Lagrangian Asymmetric-vTwin SVR in terms of pinball loss and total error metrics. The Linear Regression model has the greatest error values, which means that it has more difficulty making accurate quantile predictions than the other two models.

Table 2: Hyperparameter Tuning Results

C Value	RMSE	MAE	Pinball Loss (α=0.5)	Training Time (s)
0.1	0.300	0.215	0.110	45
1.0	0.252	0.181	0.082	56
10.0	0.265	0.195	0.095	63
100.0	0.310	0.230	0.120	72

The table displays the results of hyperparameter tweaking for different values of the regularization parameter CCC in a model. It shows how these values affect RMSE, MAE, pinball loss (α = 0.5), and training time. When CCC grows from 0.1 to 100, the RMSE and MAE first decline and reach their lowest values at C=1.0 (0.252 and 0.181, respectively). After that, they increase somewhat again at higher values of CCC. Similarly, the Pinball Loss (α = 0.5) is maximized at C=1.0C = 1.0C=1.0 (0.082), and increases for increasing values of CCC. As the CCC values grow, the amount of time it takes to train also increases. At C=0.1, it takes 45 seconds, and at C=100.0, it takes 72 seconds. This is because bigger regularization values demand more computing work. In general, C=1.0C=1.0C=1.0 offers the most effective combination of performance and training efficiency.

V. CONCLUSION

The results show that the Lagrangian Asymmetric-vTwin SVR is better than both the Standard SVR and Linear Regression models in every metric that was assessed. In particular, it regularly produces the lowest Pinball Loss values for all quantiles (α = 0.1, 0.5, 0.9), as well as the lowest RMSE and MAE values, which shows that it is more accurate than other methods when it comes to regression jobs. The hyperparameter tweaking of the SVR models shows that the optimum regularization parameter (C = 1.0) gives the best balance between prediction performance and training time, with the lowest RMSE, MAE, and Pinball Loss (α = 0.5). Furthermore, increasing the C value beyond 1.0 results in a little decrease in performance, as well as lengthier training sessions. In general, the study shows that the Lagrangian Asymmetric-vTwin SVR model is a strong method for regression problems that involve quantile predictions, especially when it is tuned with the right hyperparameters.

REFERENCES

1. S. Dang, L. Peng, J. Zhao, J. Li, and Z. Kong, "A Quantile Regression Random Forest-Based Short-Term Load Probabilistic Forecasting Method," *Energies*, vol. 15, no. 2, pp. 1–20, 2022.
2. C. Sigauke, M. M. Nemukula, and D. Maposa, "Probabilistic Hourly Load Forecasting Using Additive Quantile Regression Models," *Energies*, vol. 11, no. 9, pp. 1–21, 2018, doi: 10.3390/en11092208.

3. L. Yu, Z. Yang, and L. Tang, "Quantile estimators with orthogonal pinball loss function," *Journal of Forecasting*, vol. 37, no. 9, pp. 401–417, 2018.
4. W. Zhang, H. Quan, and D. Srinivasan, "An Improved Quantile Regression Neural Network for Probabilistic Load Forecasting," *IEEE Transactions on Smart Grid*, vol. PP, no. 9, pp. 1–1, 2018, doi: 10.1109/TSG.2018.2859749.
5. D. Gan, Y. Wang, S. Yang, and C. Kang, "Embedding Based Quantile Regression Neural Network for Probabilistic Load Forecasting," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 244–254, 2018, doi: 10.1007/s40565-018-0380-x.
6. M. Fasiolo, Y. Goude, R. Nedellec, and S. Wood, "Fast Calibrated Additive Quantile Regression," *Journal of the American Statistical Association*, vol. 116, no. 535, pp. 1–26, 2017.
7. T. Hu, D.-H. Xiang, and D.-X. Zhou, "Online learning for quantile regression and support vector regression," *Journal of Statistical Planning and Inference*, vol. 142, no. 12, pp. 3107–3122, 2012, doi: 10.1016/j.jspi.2012.06.010.
8. I. Steinwart and A. Christmann, "Estimating conditional quantiles with the help of the pinball loss," *Bernoulli*, vol. 17, no. 1, pp. 211–225, 2011, doi: 10.3150/10-BEJ267.
9. S. Zheng, "Gradient descent algorithms for quantile regression with smooth approximation," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 3, pp. 191–207, 2011, doi: 10.1007/s13042-011-0031-2.
10. G. Biau and B. Patra, "Sequential Quantile Prediction of Time Series," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1664–1674, 2011, doi: 10.1109/TIT.2011.2104610.
11. J. Park and J. Kim, "Quantile regression with an epsilon-insensitive loss in a reproducing kernel Hilbert space," *Statistics & Probability Letters*, vol. 81, no. 1, pp. 62–70, 2011, doi: 10.1016/j.spl.2010.09.019.
12. M. Somers and J. Whittaker, "Quantile regression for modelling distributions of profit and loss," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1477–1487, 2007, doi: 10.1016/j.ejor.2006.08.063.