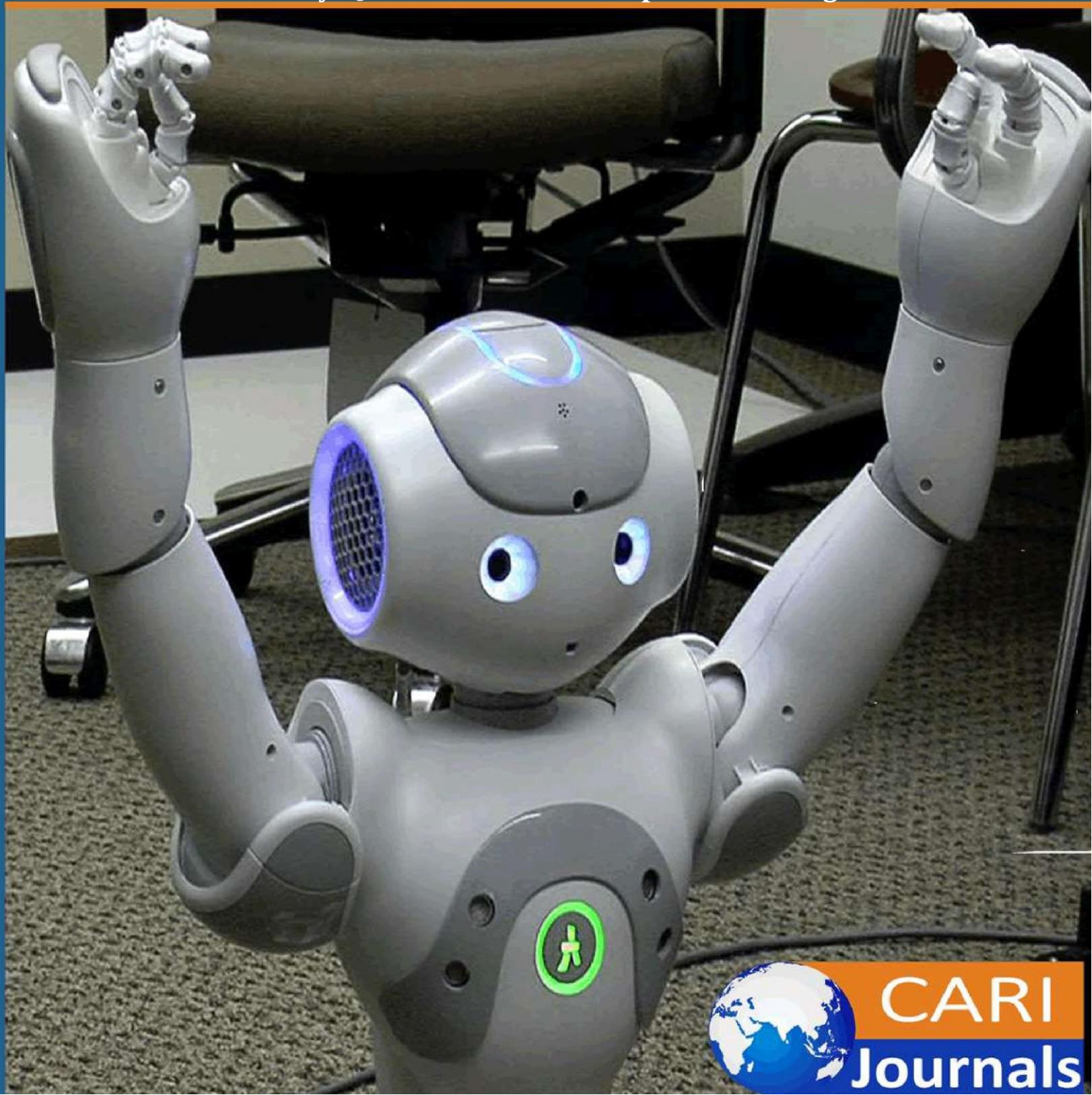
International Journal of
Computing and
Engineering
(IJCE)

Comparing Conformal and Quantile Regression for Uncertainty Quantification: An Empirical Investigation



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 5, Issue No. 5, pp. 1 - 8, 2024
www.carijournals.org

Comparing Conformal and Quantile Regression for Uncertainty Quantification: An Empirical Investigation

[1*] Bhargava Kumar, [2] Tejaswini Kumar, [3] Swapna Nadakuditi, [4] Hitesh Patel, [5] Karan Gupta

[1*, 2] Independent Researcher, Columbia Univ Alumni

[3] Sr IT BSA, Florida Blue

[4] Independent Researcher, NYU Univ Alumni

[5] Staff Data Scientist, SunPower Corporation

https://orcid.org/0009-0005-8624-2354

Accepted: 27th Mar, 2024 Received in Revised Form: 27th Apr, 2024 Published: 27th May, 2024

Abstract

Purpose: This research assesses the efficacy of conformal regression and standard quantile regression in uncertainty quantification for predictive modeling. Quantile regression estimates various quantiles within the conditional distribution, while conformal regression constructs prediction intervals with guaranteed coverage.

Methodology: By training models on multiple quantile pairs and varying error rates, the analysis evaluates each method's performance.

Findings: Results indicate consistent trends in coverage and prediction interval lengths, with no significant differences in performance. Quantile regression intervals lengthen toward the distribution tails, while conformal regression intervals lengthen with higher coverage.

Unique contribution to theory, policy and practice: On the tested dataset, both methods perform similarly, but further testing is necessary to validate these findings across diverse datasets and conditions, considering computational efficiency and implementation ease to determine the best method for specific applications.

Keywords: *Uncertainty Quantification, Machine Learning, Quantile Regression, Conformal Regression, Prediction Intervals, Error Rate, Catboost*

Introduction

In the realm of predictive modeling, providing not just single-point predictions but also reliable measures of uncertainty is crucial for informed decision-making. For instance, financial forecasting, where a point estimate of future stock prices might not adequately capture the risk associated with investment decisions. In contrast, having confidence bands around predictions enables investors to assess the potential range of outcomes, thus allowing for more robust risk management strategies.[1]

Quantile regression [2] is a sophisticated solution that offers a comprehensive view of uncertainty. It achieves this by estimating different quantiles of the conditional distribution. Technically, quantile regression estimates the conditional quantiles of the response variable based on the predictor variables. It does this by minimizing a specific loss function, such as the pinball loss, which penalizes deviations from predicted quantiles. For instance, if an 80% coverage prediction interval is required, then the 10th and 90th quantiles are computed, and corresponding intervals are formed. This approach allows quantile regression to capture varying degrees of uncertainty even for highly heteroscedastic data, and it is adaptive to local variability [1],[3-6].

Conformal regression is also an alternative to traditional regression by providing guaranteed coverage intervals for predictions [7-10]. Unlike point estimates, which offer a single value, conformal regression constructs an interval around the predicted value. This interval ensures the true value falls within it with a user-defined coverage level (e.g., 90%) – a probabilistic guarantee absent in traditional methods. The methodology achieves this by leveraging non-conformity scores, which assess the deviation of a new data point's predicted value from the model's predictions on the calibration data, augmented with the new point's true value (kept hidden during calibration). The model then ranks the nonconformity scores to find the score threshold corresponding to the required error rate and then constructs prediction intervals around the point estimate. This framework is particularly appealing due to its distribution-free nature – it doesn't require assumptions about the underlying data distribution, making it robust to various scenarios. Another positive aspect of this approach is validation by a recent scientific study which statistically validated that conformal sets improve human decision-making compared to other fixed prediction sets, highlighting their efficacy in enhancing human-AI collaboration [11]. Nonetheless, a crucial trade-off exists: higher coverage guarantees wider intervals, and vice versa. Therefore, conformal regression empowers researchers to make informed decisions by balancing these aspects based on their specific needs.

The objective of this study was to evaluate the effectiveness of conformal regression in comparison to standard quantile regression for uncertainty quantification within predictive modeling. Our investigation centered on assessing whether the conformal approach outperformed conventional quantile regression regarding both coverage and the average length of prediction intervals. Through a comprehensive analysis across various confidence intervals, we aimed to determine which method offered superior performance in accurately estimating uncertainty.

Literature Review

Quantile Regression

Quantile regression, a statistical method introduced by Koenker and Bassett in 1978 [2], has garnered considerable attention due to its capacity to provide a comprehensive understanding of conditional distribution. While traditional regression methods aim to estimate the conditional mean of the response variable, quantile regression expands upon this conception by estimating various quantiles of the conditional distribution. This approach grants valuable insights into the entire distribution of the response variable rather than merely its central tendency, rendering it particularly beneficial in circumstances where the distribution is asymmetric or heteroscedastic [1].

Quantile regression is well-suited for managing datasets with outliers or heavy-tailed distributions, as it allows for the estimation of conditional quantiles that are less influenced by extreme observations in comparison to the conditional mean. By estimating multiple quantiles, quantile regression presents a more intricate view of uncertainty and variability in the data, enabling researchers to assess the influence of varying factors across diverse segments of the distribution. Furthermore, quantile regression is inherently resilient to violations of the homoscedasticity assumption, making it applicable to a broad range of real-world datasets [4].

Quantile regression has been widely utilized in a range of fields, such as economics, finance, environmental science, healthcare, real estate and others [1],[3-4]. For instance, in finance, it is employed to estimate Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), which are crucial risk measures for portfolio management and risk assessment. In healthcare, quantile regression is used to analyze the relationship between patient characteristics and medical outcomes across different quantiles of the response variable, providing valuable insights for personalized medicine and healthcare policy.

Overall, quantile regression presents a flexible and powerful framework for modeling conditional distributions and quantifying uncertainty in predictive modeling tasks [1-6]. Its capacity to capture the heterogeneity and asymmetry in the data makes it a valuable tool for researchers and practitioners seeking a deeper understanding of the relationships between variables and the variability in their outcomes.

Conformal Prediction

Conformal prediction represents a paradigm shift in the realm of predictive modeling by offering a systematic approach to uncertainty quantification. Traditionally, predictive modeling focused on providing point estimates, such as mean or median predictions, without quantifying the associated uncertainty. However, in many real-world applications, decision-makers require not only predictions but also reliable measures of uncertainty to assess the risk associated with their decisions [10].

Conformal prediction addresses this need by providing prediction intervals that come with statistical guarantees of coverage probability. Unlike traditional point prediction methods, which

provide single-point estimates, conformal prediction constructs prediction intervals that contain the true value with a specified confidence level. This framework offers a probabilistic guarantee that the true value falls within the interval, making it particularly valuable when uncertainty quantification is essential for decision-making under uncertainty [10-11].

At the core of conformal prediction lies the notion of "conformity," quantifying the extent to which a new data point aligns with the patterns observed in the calibration data. The framework computes a non-conformity score for each prediction, reflecting the deviation of the new data point's predicted value from those of the calibration data. Low non-conformity scores correspond to high conformity, indicating strong confidence in the prediction, whereas high non-conformity scores signify lower conformity and increased uncertainty. Subsequently, the model ranks these non-conformity scores and constructs prediction intervals around point predictions (e.g., median or mean). These intervals are designed to achieve a pre-defined coverage level, providing a probabilistic guarantee that the true value falls within the interval [7],[10].

One of the key advantages of conformal prediction is its distribution-free nature. Unlike parametric methods that rely on assumptions about the underlying data distribution, conformal prediction does not make any distributional assumptions, making it robust to deviations from model assumptions and data distributional changes. This flexibility makes conformal prediction applicable to a wide range of modeling tasks and data types, including non-standard and complex data structures [7],[10-11].

Conformal prediction has been successfully applied in various domains, including classification, regression, anomaly detection, and time series forecasting. In regression tasks, conformal prediction constructs prediction intervals around point estimates, providing a measure of uncertainty in the predictions. This uncertainty quantification is crucial for decision-making in fields such as finance, healthcare, and environmental science, where accurate risk assessment and uncertainty estimation are paramount [10].

Overall, conformal prediction offers a principled framework for uncertainty quantification in predictive modeling, providing statistically valid prediction intervals with guaranteed coverage probabilities. Its distribution-free nature, flexibility, and ability to provide probabilistic guarantees make it a valuable tool for researchers and practitioners seeking reliable uncertainty estimates in their predictive modeling tasks [7],[9-10].

Experiment

Experiment Setup

In our experiments, we conducted both quantile regression and conformal regression analyses on the dataset, training the models for multiple pairs of quantiles: specifically, 5th and 95th, 10th and 80th, and 15th and 75th percentiles. For each pair of quantiles, we aimed to obtain prediction intervals corresponding to regions bound by them. Additionally, in the case of the conformal regression model, we trained the model against varying error rates corresponding to those of quantile pairs, which were 10%, 20%, and 30% respectively. This approach allowed us to evaluate

the performance of both regression techniques comprehensively. Subsequently, we compared the empirically observed coverage and the average length of prediction intervals obtained from both approaches. Through this comparative analysis, we aimed to gain insights into the effectiveness and suitability of quantile regression and conformal regression methods for providing reliable estimates and prediction uncertainties in regression tasks.

Dataset

The dataset used in this study is the Concrete Compressive Strength dataset [12] sourced from the UCI Machine Learning Repository. It contains data on the compressive strength of concrete samples, vital in civil engineering and construction. The dataset contains 1030 samples and includes factors like cement, slag, fly ash, water, superplasticizer, and aggregates, alongside age and compressive strength measurements. More detailed descriptions of features can be found in the accompanying figure (Figure 1), which was taken from their website. This dataset enables the exploration of relationships between input variables and concrete strength, aiding predictive modeling and optimization in concrete technology and structural engineering.

Variable Name	Role	Type	Description	Units	Missing Values
Cement	Feature	Continuous		kg/m^3	no
Blast Furnace Slag	Feature	Integer		kg/m^3	no
Fly Ash	Feature	Continuous		kg/m^3	no
Water	Feature	Continuous		kg/m^3	no
Superplasticizer	Feature	Continuous		kg/m^3	no
Coarse Aggregate	Feature	Continuous		kg/m^3	no
Fine Aggregate	Feature	Continuous		kg/m^3	no
Age	Feature	Integer		day	no
Concrete compressive strength	Target	Continuous		MPa	no

Figure 1: Detailed description of features and target of the concrete dataset from UCI Machine Learning Repository

Models

For quantile regression, we utilize CatBoost [13], benefiting from its robustness to outliers and nonlinear relationships. CatBoost's customizable loss functions enable the estimation of multiple quantiles, while its handling of missing values ensures model efficiency. Additionally, CatBoost's built-in features like categorical encoding and early stopping enhance model performance and

interpretability. Leveraging these capabilities, we aim to accurately estimate quantiles and gain insights into the distribution of the target variable

For conformal regression, we utilize the conformal prediction framework to derive prediction intervals around CatBoost's 50th quantile output. Employing the **nonconformist** library in Python, we initialize an Inductive Conformal Predictor (ICP) with the trained CatBoost regressor. This facilitates the construction of prediction intervals, offering a measure of confidence in the predictions. Through this approach, we obtain dependable estimates of prediction uncertainty, which assist in informed decision-making for regression tasks.

Results

The results of our analysis comparing quantile regression and conformal regression for uncertainty quantification, as suggested in Table 1, demonstrate consistent trends in coverage and prediction interval lengths across different quantile pairs or error rates. Notably, the coverage obtained closely aligns with the theoretically expected coverage, indicating the reliability of both regression methods in estimating uncertainty [1], [4], [9], [10]. This suggests that practitioners may not go wrong choosing either approach, as both methods result in the expected coverage while maintaining similar lengths of prediction intervals, indicating their comparable performance in uncertainty estimation.

Also, from the results we observe that both quantile regression and conformal regression exhibit varying levels of coverage and average length of prediction intervals across different quantile pairs or error rates. In the case of quantile regression, as the quantiles move toward the tail of the distribution, the average length of the prediction intervals tends to increase. This trend is consistent with the essence of quantile regression. As the quantiles forming the coverage interval extend towards the distribution's tail, wider prediction intervals become essential to encompass the increased variability in the data [1], [4]. Similarly, for conformal regression, as the error rates increase, indicating a higher tolerance for deviations from conformity, the prediction intervals' average length may tend to decrease. This aligns with the essence of conformal regression, where accommodating smaller coverage guarantee leads to smaller prediction intervals [1], [10].

Table 1: Empirical coverage and average length of prediction intervals for different quantiles, along with corresponding error rates for both the quantile and conformalized versions of the CatBoost model for the concrete dataset.

	Quantile Regression			**Conformal Regression**		
Quantiles/error rate	**[5%, 95%]**	**[10%, 90%]**	**[15%, 85%]**	**α = 10%**	**α = 20%**	**α = 30%**
Coverage	89.7	81.6	69.6	90.3	80.2	70.2
Avg Length	13.19	8.37	6.05	12.81	8.27	6.09

Conclusion

Our study aimed to evaluate the effectiveness of conformal regression compared to standard quantile regression for uncertainty quantification in predictive modeling. Through a comprehensive analysis, we found no significant difference between the two approaches in terms of coverage or average length of prediction intervals, suggesting that one approach may not inherently be superior to the other in real-world scenarios. However, the conformal approach offers a notable advantage of statistically guaranteed coverage, absent in quantile regression.

Recommendations

The results of our study highlight the significance of incorporating diverse datasets and employing statistical validation techniques to reinforce inferences. Additional research is necessary to delve into the intricacies of performance and identify the scenarios in which each method outperforms the other. Furthermore, it is imperative for researchers to evaluate the computational efficiency and practicality of both conformal regression and quantile regression, as this will enable practitioners to make well-informed decisions about applying uncertainty quantification in predictive modeling tasks.

References

[1] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," *Advances in neural information processing systems*, vol. 32, 2019.

[2] R. Koenker and G. Bassett, "Regression Quantiles," Econometrica, vol. 46, no. 1, p. 33, Jan. 1978, doi: https://doi.org/10.2307/1913643.

[3] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, Art. no. 4, 2001.

[4] N. Meinshausen and G. Ridgeway, "Quantile regression forests.," *Journal of machine learning research*, vol. 7, Art. no. 6, 2006.

[5] I. Steinwart and A. Christmann, "Estimating conditional quantiles with the help of the pinball loss," 2011.

[6] I. Takeuchi, Q. Le, T. Sears, and A. Smola, "Nonparametric quantile estimation," MIT Press, 2006.

[7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*, vol. 29. Springer, 2005.

[8] V. Vovk, I. Nouretdinov, and A. Gammerman, "On-line predictive linear regression," JSTOR, 2009.

[9] J. Lei, J. Robins, and L. Wasserman, "Distribution-free prediction sets," *Journal of the American Statistical Association*, vol. 108, Art. no. 501, 2013.

[10] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and Distribution-Free uncertainty quantification," *arXiv.org*, Jul. 15, 2021. https://arxiv.org/abs/2107.07511

[11] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis, "Conformal prediction sets improve human decision making," *arXiv.org*, Jan. 24, 2024. https://arxiv.org/abs/2401.13744

[12] I.-C. Yeh, "Concrete Compressive strength." UCI Machine Learning Repository, 2007. doi: 10.24432/C5PK67.

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[14] Donlnz, "GitHub - donlnz/nonconformist: Python implementation of the conformal prediction framework.," *GitHub*. https://github.com/donlnz/nonconformist

©2023 by the Authors. This Article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/)