Quantile Regression Analysis of House Price Determinants

Mayank Kumar, *110062284, MSc AI Specialization, kumar48@uwindsor.ca,*
Subham Gupta, *110093436, MAC AI Specialization, gupta2d@uwindsor.ca,*
and Rayhaan Pirani, *110046715, MSc AI Specialization, piranir@uwindsor.ca*

Abstract—Standard supervised regression techniques assume that the relationship between the features and target variable is consistent across the data distribution, which may not always be accurate. Quantile regression is a valuable technique for identifying how different variables relate to the target variable at various distribution points. This approach can be particularly beneficial in scenarios where the relationship between the features and target variable is inconsistent across the data distribution, which is often the case in real-world datasets. This study employs quantile regression techniques to identify how different housing characteristics relate to sale prices at various points of the sale price distribution. We consider the statistical significance of the selected features and how it behaves across different sale price distribution. Our results confirm that different housing characteristics exhibit varying significance levels at different quantiles of the sale price distribution. These results provide practitioners in the real estate industry with a more comprehensive understanding of how various features impact the overall sale price of a property. By identifying the most significant features at different points of the sale price distribution, practitioners can make more informed decisions about pricing and marketing strategies, ultimately leading to improved business outcomes.

1 Introduction

ONE of the fundamental techniques in supervised regression problems is to study the relationship between features and the target variable. The standard approach assumes that this relationship is constant across the entire range of the target variable. The ordinary least squares (OLS) method is an example of such a standard approach.

However, the relationship may vary across the range of target variables, and hedonic regression analysis may not provide the best interpretation of the underlying relationship. Quantile regression (QR) models the relationship between features and the target variable across different quantiles. We shall demonstrate this with an example of household food expenditure. Household food expenditure varies even more widely for high-income households than low-income ones, as seen in **Fig. 1**.

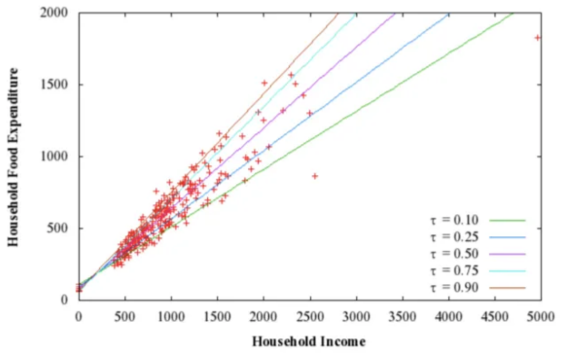

Fig. 1. Household income versus food expenditure [1]

In linear regression or ordinary least squares (OLS), we assume that the relationship between our set of input variables X and our output label Y can be modeled by a linear function by minimizing the sum of squared errors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{1}$$

$$\mathrm{L}(.) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 = \sum_{i=1}^{n} (e_i)^2 \tag{2}$$

Median regression minimizes the absolute deviation to obtain the estimator. If X is symmetric, the mean and median will be approximately the same, and the estimator will be the same as that of OLS. Quantile regression uses Least-Absolute-Deviation (LAD) to obtain the estimators.

$$\mathrm{L}(.) = \sum_{i=1}^{n} |e_i| \tag{3}$$

LAD regression does not have an analytical solving method like an OLS, such as linear programming or the simplex method. It searches for an estimator that satisfies the following requirement (linear programming):

$$\begin{gathered} \min \sum_{i=1}^{n} t_i \\ -t_i \le y_i - x_i\beta \le t_i \end{gathered} \tag{4}$$

LAD can be extended to quantile regression. Quantile regression minimizes a sum that gives asymmetric penalties.

$$\mathrm{L} - \mathrm{q}(.) = q \sum_{i=1}^{n} |e_i| + (1-q) \sum_{i=1}^{n} |e_i| \tag{5}$$

Where $q \in (0,1)$ for the q^{th} quantile. When $q = 0.50$, the quantile regression collapses to OLS.

2 Problem Statement

In this section, we define the problem statement, the motivation for the problem statement and the justification for our approach to solving the problem.

2.1 Problem Definition

Given a housing prediction dataset (DS) with explanatory features (FS), train a $q + 1$ model – q Quantile Regression Model and 1 OLS model. Investigate and compare these models to provide an analysis of House Price Determinants.

Here,
$DS =$ Ames Housing Dataset
$FS =$ 79 explanatory features in DS
$q = [0.1, 0.5, 0.9]$

2.2 Motivation

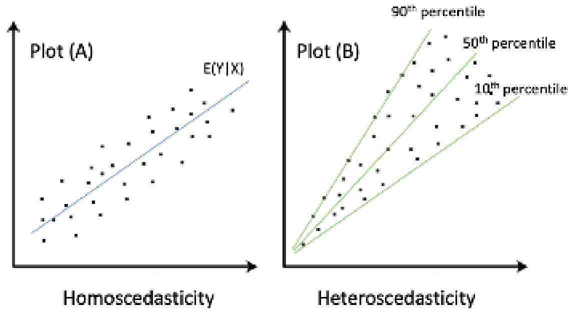

Fig. 2. Homoskedasticity vs. Heteroscedasticity [2]

In **Fig. 2**, Plot (A) shows the case where the variance of Y stays the same, and in Plot (B), the variance of Y increases as X increases. In econometrics, Plot (A) is called homoskedasticity, in which the variance of the residual term remains constant. Plot (B) is called heteroscedasticity, where the variance varies widely. Most real-world data shows heteroscedasticity, i.e., the data is closer to the representation in Plot (B) than in Plot (A).

Quantile regression models may thus help better investigate and validate whether house characteristics are priced the same or different across a given distribution of house prices. This concept is a solid motivator for our study to use quantile regression methods over OLS.

2.3 Justification

The use of quantile regression techniques is justified in this research study due to the inconsistent relationship between features and the target variable across the data distribution, which is often the case in real-world datasets. The standard supervised regression techniques, such as OLS, assume that this relationship is constant across the entire range of the target variable, which may not be accurate.

Quantile regression models the relationship between features and the target variable across different quantiles. It is a valuable technique for identifying how different variables relate to the target variable at different distribution points. This approach is advantageous in scenarios where the relationship between the predictors and the target variable is inconsistent across the data distribution.

Furthermore, heteroscedasticity is prevalent in most real-world datasets, where the variance of the residual term varies widely. By identifying the most significant features at different points of the sale price distribution, practitioners in the real estate industry can make more informed decisions about pricing and marketing strategies, ultimately leading to improved business outcomes. This may be achieved using more feature-enriched datasets, such as the Ames Housing dataset [3], with 79 explanatory features. We then compare the findings of the quantile regression analysis with those of ordinary least squares for house pricing determinants.

3 Related Work

Sirmans et al. in 2005 [4] discuss using hedonic regression analysis to estimate the marginal contribution of various characteristics that affect house prices. Their paper reviews 125 empirical studies and finds that specific characteristics, such as a slanted roof, sprinkler system, and gated community, positively affect the selling price. In contrast, others, such as living in an earthquake zone, proximity to a landfill, and properties that require flood insurance, negatively affect the selling price. The paper emphasizes the importance of considering multiple characteristics when estimating the value of a house. They find that the studies often disagree on the magnitude and direction of the effect of certain characteristics.

Malpezzi et al. in 2003 [5] provide a review of the theoretical development of hedonic pricing models, which are used to estimate the value of individual characteristics of a house. The article emphasizes that the hedonic model is necessary due to the heterogeneity of the housing stock and the consumers. Each house has different characteristics that may be valued differently by different consumers. The article highlights the importance of considering these factors when estimating the value of a house.

Newsome and Zietz in 1992 [6] discuss using multiple regression analysis (MRA) by appraisers to adjust comparable sales in the market sales comparison approach. The authors highlight the issue of heteroscedasticity, where housing characteristics may not be valued the same across a given distribution of housing prices. The authors propose solving this problem by developing separate MRA models for houses in different price ranges. This approach can help minimize the distortions caused by heteroscedasticity and can also be applied to commercial properties.

4 Methodology

4.1 Material and Dataset

The material and data used in this study were sourced from the Ames Housing dataset [3]. The dataset comprises 1460 records and 79 explanatory variables that describe almost every aspect of residential homes in Ames, Iowa. The dataset provides features that capture various housing characteristics, including the property size, number of bedrooms and bathrooms, amenities such as garages and pools, and the overall property condition. The richness of the data allows for a thorough investigation of the relationship between the different housing characteristics and property sale prices.

4.2 Proposed Models

This study employs two regression techniques to investigate the relationship between housing characteristics and sale prices: Ordinary Least Squares (OLS) and Quantile Regression (QR).

OLS is a widely used technique that estimates the parameters of a linear regression model by minimizing the sum of squared residuals. It assumes that the relationship between the features and the target variable is consistent across the data distribution and aims to find the line of best fit that minimizes the distance between the observed data points and the predicted values.

On the other hand, QR is a technique that models the relationship between the features and the target variable across different quantiles. The quantile loss function used in QR differs from the least squares loss function in OLS. The quantile loss function is defined as follows:

$$L = \begin{cases} q(y - X\theta), \text{ if } y - X\theta \geq 0 \\ (q-1)(y - X\theta), \text{ if } y - X\theta < 0 \end{cases} \tag{6}$$

Where L is the loss function, q is the quantile level, y is the observed value, and $X\theta$ is the predicted value. The quantile loss function penalizes positive and negative errors differently depending on the specified quantile level. Specifically, more negative errors (over-predicting) are penalized more when we specify higher quantiles and more positive errors (under-predicting) are penalized more for lower quantiles.

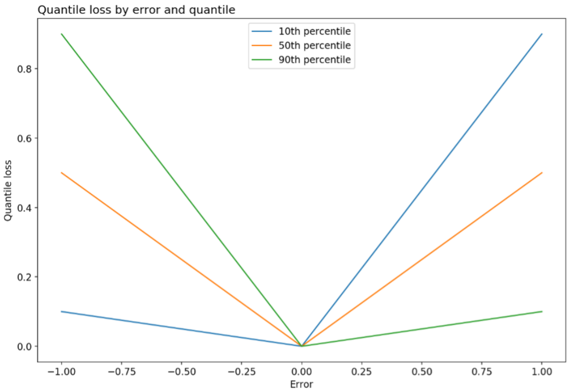

Fig. 3. QR Loss plot with $q = 0.1, 0.5, 0.9$ for dummy data

By employing both OLS and QR in our analysis, we can investigate how different housing characteristics relate to sale prices at different points of the sale price distribution. We can thus determine which features exhibit varying significance levels at different quantiles of the sale price distribution.

4.3 Condition and Assumption

The study investigates the relationship between determinants and house prices based on the variation in house prices, such as – high and low-price range houses. The study does not consider any external or other relationship, such as the spatial location for the same.

4.4 Formal Complexity

Linear programming approaches are commonly used to solve the Quantile LAD regression problem. The computational complexity of these approaches depends on the number of observations and independent variables. Specifically, the computational complexity of solving a linear program with n variables and m constraints is typically of order $O(n^3 + nm^2)$, which can become quite large for large datasets with many variables. As such, it may be necessary to explore alternative methods for solving the Quantile LAD regression problem for large datasets, such as gradient-based approaches or heuristic methods.

5 Computational Experiments

5.1 Experiments

We trained three Quantile Regression models for percentiles $q = [0.1, 0.5, 0.9]$ and one OLS model in our computational experiments. Before training the models, we conducted exploratory data analysis and handled missing values. We used vectorization with one-hot encoding and normalization with Min-Max Scalar to prepare the data for modelling. For this purpose, we used feature selection with univariate and forward selection based on Akaike Information Criterion (AIC) and multi-collinearity techniques.

We then conducted a train-test split of the data and trained the models using the specified techniques and multiple visualization techniques. Overall, these experiments helped us to identify which variables are most strongly associated with the sale price and how the strength of these relationships varies across the price range.

5.2 Metrics

We have reported Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and for all the models. Also, we have reported the Quantile Loss for the Quantile Regression models. Additionally, we investigated how different housing price determinants/features are associated with sale price distribution, particularly for high-price range houses versus low-price range houses. We visualized these relationships using scatterplots and barplot to gain insights into how the models capture the underlying relationships between the features and target variables at different quantiles. We also examined the stability of the models across different quantiles by comparing the coefficients and significance levels of the variables.

5.3 Implementation Details

The project was implemented on Google Colab using Python and various toolkits and libraries such as statsmodels, Scikit-learn and Seaborn. The code was managed using the Git version control system. These tools and technologies were chosen for their versatility and ease of use in implementing machine learning algorithms and performing data analysis tasks.

After collecting the dataset, we performed exploratory data analysis (EDA) on various data types and handled any missing values. Our dataset consisted of continuous, discrete, categorical, and year data types. For continuous

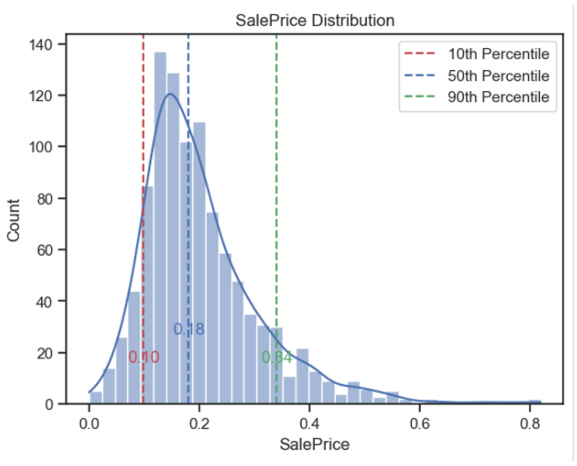

Fig. 4. SalePrice distribution – overall data (normalized)

data types, such as LotFrontage and MasVnrArea, which had missing values, we replaced these with the median value from the neighbourhood of the housing record. There were no missing values for discrete data types. For year data types, such as GarageYrBlt (garage year built), missing values were filled with values from the year built for the house. Lastly, we had 16 columns with null values for categorical data types, which we replaced with the string 'None'.

The plot in **Fig. 4** shows the sale price distribution and marks the 0.1, 0.5, and 0.9 values.

5.3.1 Feature Selection

Our study used various feature selection techniques to identify the most relevant variables for the Quantile LAD Regression model. We began by applying filter methods, including basic methods to remove constant and quasi-constant features, which are variables that display the same value for all or a great majority of the observations of the dataset. We also utilized the correlation matrix with heatmap visualization to identify variables that exhibit a linear relationship with the target variable but are uncorrelated among themselves.

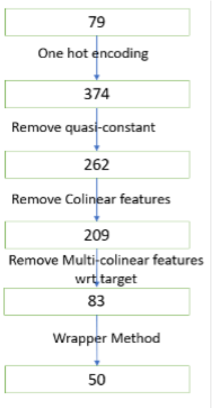

Fig. 5. Feature dimension across various steps

Furthermore, we employed the wrapper forward selection method, using Akaike Information Criterion (AIC) as the performance indicator. This method selects the optimal subset of variables by progressively adding features that lead to the lowest AIC value. This approach allows for a more fine-grained selection of variables that significantly impact the target variable, thereby improving the accuracy and efficiency of the model. **Fig. 5** shows the dimension size starting from the initial dimension of 79 features to the final dimension of 50 after completing the above-mentioned feature selection techniques.

Overall, our feature selection process involved a combination of basic filtering techniques, correlation analysis, and wrapper methods to identify the most relevant variables for our Quantile LAD Regression model.

5.3.2 Data Split

To evaluate our models' performance, we partitioned our dataset into two subsets: a training set and a testing set. The split was performed using a 70-30 ratio; 70% of the data was used for training the models, and 30% was used for testing the trained models. This approach allowed us to measure the generalization ability of the models and gain interpretability using various quantile models and whether the modelled relation holds on unseen test data.

5.3.3 Model Train and Test Metrics

We reported multiple median metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). We have also reported Quantile Loss. Below we showcase these metrics in **Table 1** and **Table 2** and plot the same metrics in **Fig. 6** and **Fig. 7**. We have evaluated these for both train and test split.

TABLE 1
Prediction metric report - train set

Model	**MAE**	**MSE**	**RMSE**	**Quantile Loss**
0.1 – QR	0.049	0.006	0.075	**0.006**
0.5 – QR	**0.029**	**0.002**	0.046	0.015
0.9 – QR	0.054	0.006	0.0075	0.007
OLS	0.031	0.002	**0.044**	–

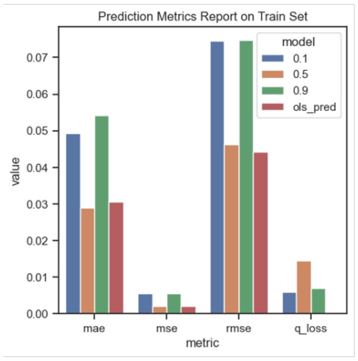

Fig. 6. Prediction metric report - train set

We see the model report similar metrics in both train and test fold which is a valuable estimator to show that our models are not over-train, and the subsequent results discussed hold great insight.

TABLE 2
Prediction metric report - test set

Model	MAE	MSE	RMSE	Quantile Loss
0.1 – QR	0.054	0.008	0.091	**0.007**
0.5 – QR	**0.033**	**0.003**	**0.051**	0.016
0.9 – QR	0.056	0.005	0.0072	0.009
OLS	0.035	0.003	0.0054	–

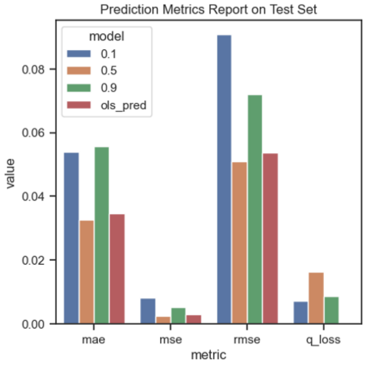

Fig. 7. Prediction metric report - test set

5.4 Results

We investigate the top 10 features for all 4 models and see if the relationship between the house price determinants is consistent with the training data. Below, we present a bar graph for the top 10 features for the 0.1 and 0.9 quantile regression models.

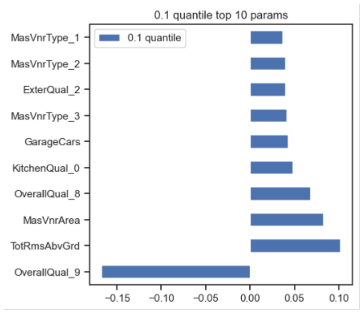

Fig. 8. 0.1 quantile top 10 features

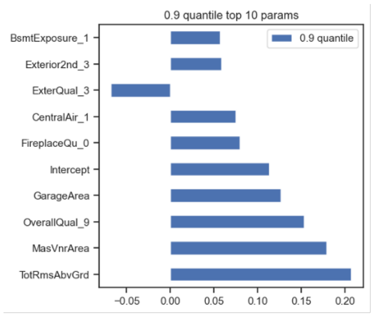

Fig. 9. 0.9 quantile top 10 features

From **Fig. 8** and **Fig. 9**, we can see that a few features, such as KitchenQual, play a vital role in determining house prices for low-range houses, which is not necessarily an essential criterion for high-price-range houses. Alternatively, feature such as BsmntExposure, FireplaceQu plays a significant role in high-price range houses. We can also see that GarageArea/GarageCars plays a crucial role in both the quantile model. Only 30% of the top 10 features - OverallQual_9, TotRmsAbvGrd, MasVnrArea are consistent house determinant features across all 3 quantiles, emphasizing the necessity of the QR approach to investigate the relationship between target and feature variables better.

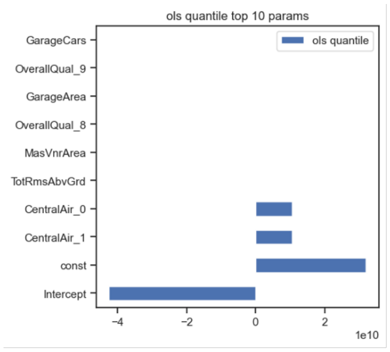

Fig. 10. OLS top 10 features

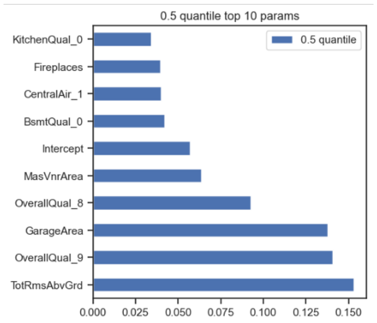

Fig. 11. 0.5 quantile top 10 features

We also plot the top 10 features for OLS and 0.5 quantile regression - **Fig. 10** and **Fig. 11**.

CentralAir is a feature which occurs across OLS and all QR model and signifies a vital characteristic to be considered as a critical housing price determinant.

Next, we also tried to investigate how the coefficient of the features varies across different quantiles for QR and OLS. **Fig. 12** shows the coefficient and the confidence interval. We can see that QR have a different coefficient value for each quantile, illustrating that QR can model different relation between feature and target variable.

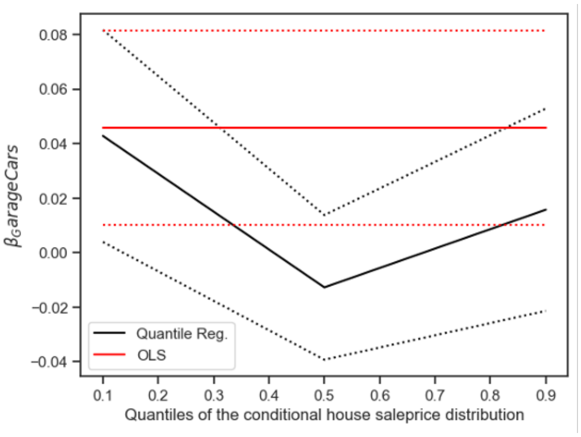

Fig. 12. GarageCars feature coefficient and interval

Lastly, we plotted the actual vs predicted plot in **Fig. 13**. The plot shows that 0.9 QR models always try to over-predict, while 0.1 QR models try to under-predict. The 0.5 and OLS models model the median/mean relationship and are between the 0.1 and 0.9 QR models.

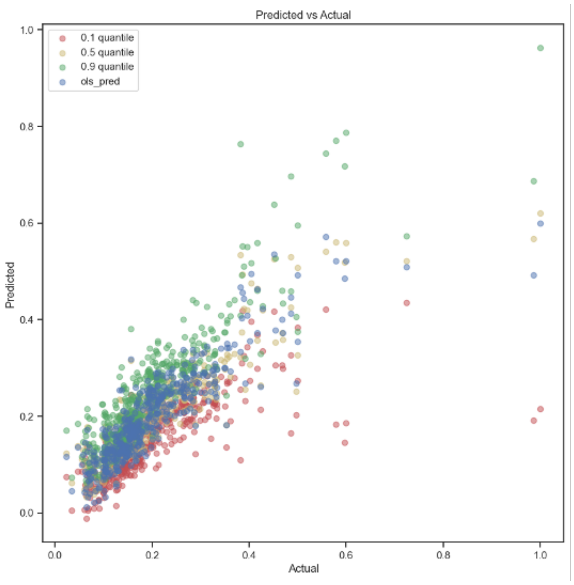

Fig. 13. Actual vs. predicted - all models

5.5 Discussion

The experimentation conducted in this study has yielded insightful results regarding the valuation of housing price determinants and features in relation to the SalePrice distribution. It was observed that different housing features are valued differently depending on the target SalePrice distribution. These findings highlight the importance of carefully considering the distribution of the target variable when modeling the housing market. This can be particularly relevant for real-world applications, where making accurate predictions on the value of a property is of utmost importance. Practitioners can utilize the results of this study to make more informed decisions while valuing properties and provide valuable insights for further research in the field of housing market modeling.

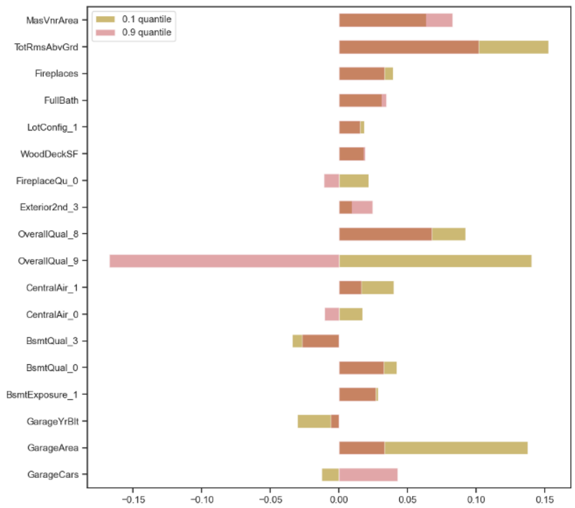

Fig. 14. Feature coefficient - 0.1 vs 0.9 QR models

Fig. 14 shows the coefficient of multiple features for which the relationship between the features and target variable is modelled differently in terms of – direction or magnitude. Many features, such as OverallQual, share opposing effects on housing prices in lower range versus higher range houses.

6 Conclusion

6.1 Summary

In summary, our experimentations using quantile regression have shown that it is a robust approach to handle outliers and provides a more inclusive study of the relationship between housing price and its determinants. Our analysis has also revealed that the valuation of housing price determinants/features varies across the target - SalePrice Distribution. Furthermore, **Fig. 15** demonstrates the housing data and the predicted interval, which shows that most observations lie within the interval. This indicates the effectiveness of our approach in predicting the housing price with a certain level of confidence.

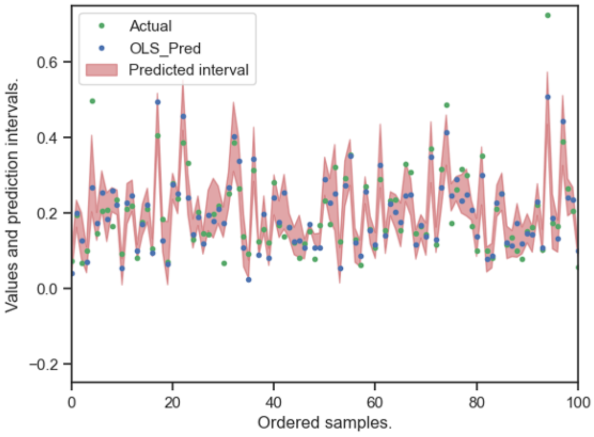

Fig. 15. Actual, predicted and interval using QR

To conclude, our study aimed to identify the most significant determinants of housing prices and their relationships

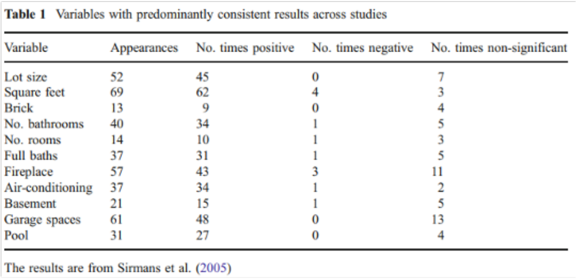

Table 1 Variables with predominantly consistent results across studies

Variable	Appearances	No. times positive	No. times negative	No. times non-significant
Lot size	52	45	0	7
Square feet	69	62	4	3
Brick	13	9	0	4
No. bathrooms	40	34	1	5
No. rooms	14	10	1	3
Full baths	37	31	1	5
Fireplace	57	43	3	11
Air-conditioning	37	34	1	2
Basement	21	15	1	5
Garage spaces	61	48	0	13
Pool	31	27	0	4

The results are from Sirmans et al. (2005)

Fig. 16. Table from Sirmans et al. [4] – consistent variable

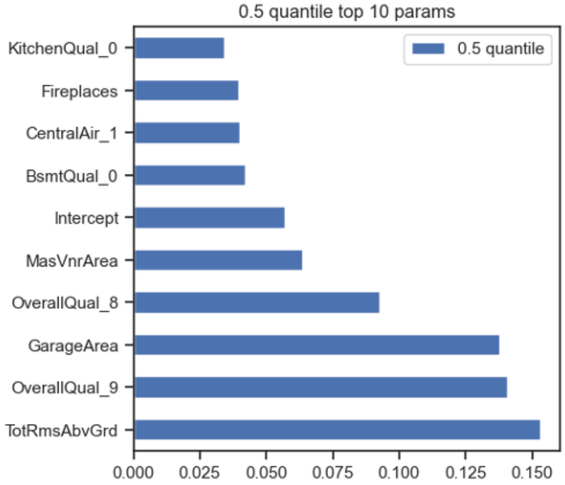

Fig. 17. 0.5 top 10 parameters

using quantile regression. Our findings are consistent with previous literature on this topic, as we identified several key features that contribute significantly to housing prices. Additionally, our study highlights the importance of quantile regression in analyzing the relationship between housing determinants and prices, providing a more inclusive and robust approach to modeling and analyzing the data. The top 10 features identified in our study also show a high degree of consistency with the features identified by Sirmans et al. (2005) [4] – [No. rooms, Fireplace, Air-conditioning, Basement, Garage spaces]. Overall, our study provides valuable insights in the housing market and can assist in better predicting and understanding housing prices.

6.2 Further Research

To further improve the model's accuracy, future research can explore other quantile-based approaches, such as tree-based methods and deep learning methods, which can handle the curse of dimensionality and feature selection more effectively. Additionally, exploring other data types, such as geospatial and weather data, can provide more insight into the housing price determinants. Furthermore, examining the impact of external factors, such as economic conditions and policy changes on the housing market, can lead to a more comprehensive understanding of the relationship between housing price and its determinants.

References

[1] Koenker, R. (2005). *Quantile Regression (Econometric Society Monographs)*. Cambridge: Cambridge University Press. 10.1017/CBO9780511754098.

[2] Chauhan, P. (2019, August 1). *A Tutorial on Quantile Regression, Quantile Random Forests, and Quantile GBM*. DataMan in AI. https://medium.com/dataman-in-ai/a-tutorial-on-quantile-regression-quantile-random-forests-and-quantile-gbm-d3c651af7516.

[3] De Cock, D. (2011). *Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project*. Journal of Statistics Education, 19(3).

[4] Sirmans, G., Macpherson, D., Zietz, E. (2005). *The Composition of Hedonic Pricing Models*. Journal of Real Estate Literature. 13. 3-43. 10.1080/10835547.2005.12090154.

[5] Turner, B., Malpezzi, S. (2003). *A Review of Empirical Evidence on the Costs and Benefits of Rent Control*. Swedish Economic Policy Review, 10, 11-56.

[6] Newsome, B. A., Zietz, J. (1992). *Adjusting comparable sales using multiple regression analysis - The need for segmentation*. Appraisal Journal, 60(1), 129-136.