Kernel Density Based Linear Regression

*Roshan Niranjan Kalpavruksha*

*Pace University*

[*rkalpavruksha@yahoo.com*](mailto:rkalpavruksha@yahoo.com)

***Abstract***

***Kernel Density-Based Linear Regression (KDLR) is a sophisticated regression technique that integrates the principles of kernel density estimation with traditional linear regression. Traditional regression methods often fall short in scenarios involving noise, outliers, or datasets with complex patterns. By dynamically weighting data points based on their density, KDLR addresses these limitations, prioritizing densely clustered data points and minimizing the influence of outliers. This report delves into the mathematical foundations of KDLR, highlights its advantages over traditional methods, and explores its applications across domains such as finance, healthcare, and environmental modelling. Experimental evaluations demonstrate that KDLR not only enhances robustness and accuracy but also effectively captures subtle data patterns, making it a promising alternative for real-world predictive tasks.***

***Keywords: Kernel Density Estimation (KDE), Weighted Linear Regression, Outlier Resistance, Data Density Analysis, Robust Regression Techniques, Pattern Recognition in Noisy Data***

**INTRODUCTION**

Linear regression has been a cornerstone of statistical modelling and machine learning for decades due to its simplicity and interpretability. Its underlying assumption that all data points contribute equally to the regression model works well for clean and uniformly distributed datasets. However, in real-world applications, datasets are rarely ideal. They often contain:

1. Outliers: Extreme values that disproportionately influence the model.
2. Noise: Random variations or errors that obscure meaningful patterns.
3. Heterogeneity: Dense clusters of data mixed with sparse, less-representative regions.

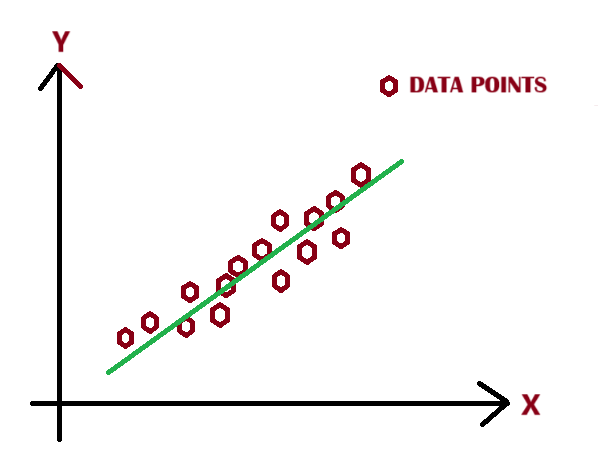


Fig 1. Linear Regression without Outliers

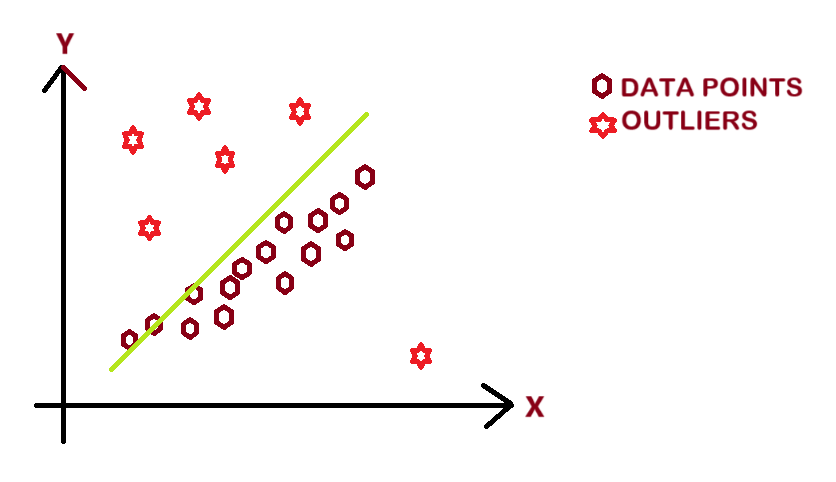


Fig 2. Linear Regression with Outliers

Motivating Questions:-

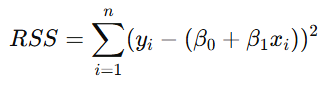
These challenges raise fundamental questions:

* How can regression models better reflect the true patterns in a dataset?
* Can data density be leveraged to improve the robustness and accuracy of regression models?

Kernel Density-Based Linear Regression (KDLR) offers an innovative solution by incorporating density-based weighting into the regression process. By assigning higher weights to densely clustered data points and lower weights to sparse or noisy regions, KDLR aligns model training with the underlying data distribution.

**LINEAR REGRESSION**

Linear regression is a supervised learning method used to model the relationship between a dependent variable Y and one or more independent variables X. The goal is to find a line (or hyperplane) that minimizes the residual sum of squares (RSS):



Where β0 is the intercept and β1 is the slope (for simple linear regression).

Strengths:-

* Interpretability: Coefficients β0 and β1 provide insights into the relationship between variables.
* Computational Efficiency: Solved efficiently using closed-form solutions or gradient descent.

Limitations:-

1. Equal Weighting: All data points are treated equally, regardless of their representativeness.
2. Sensitivity to Outliers: Large deviations can skew the fitted line significantly.
3. Inability to Model Nonlinear Relationships: Assumes a linear relationship, missing subtle patterns in complex data.

These limitations necessitate the exploration of alternative methods like Kernel Density-Based Linear Regression.

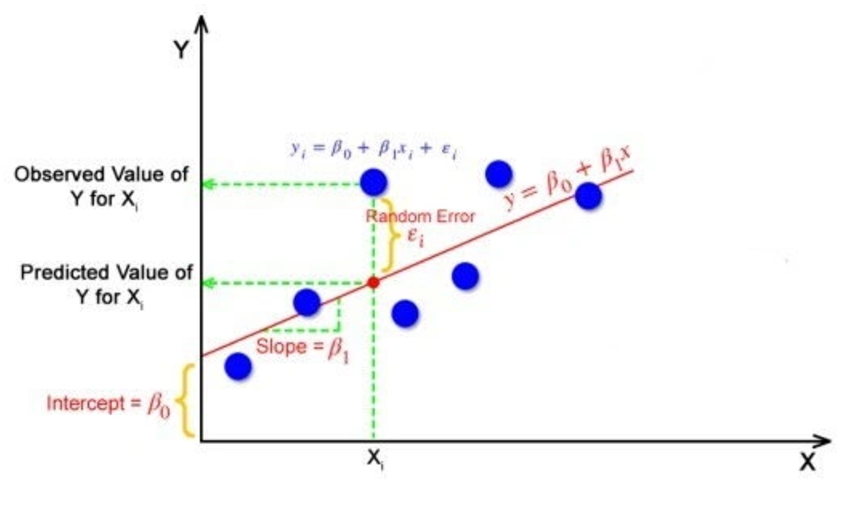


Fig 3. Linear Regression representation

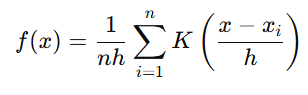
**KERNEL DENSITY ESTIMATION**

Overview:-

Kernel Density Estimation (KDE) is a non-parametric method for estimating the probability density function of a dataset. Unlike histograms, which partition data into discrete bins, KDE provides a continuous estimate by smoothing data points using a kernel function.

Mathematical Formulation:-

For a dataset X={x1, x2, …, xn}, the KDE at a point x is given by:



Where:

* K is the kernel function (e.g., Gaussian, Uniform)
* h is the bandwidth, controlling the smoothing level.

Properties:-

1. Data Density Insight: Provides a smooth estimate of data density.
2. Outlier Identification: Sparse regions correspond to lower density, helping distinguish anomalies.

KDE’s ability to identify dense and sparse regions is central to KDLR, where it informs the weighting mechanism.

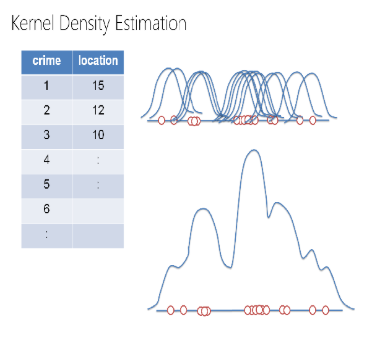


Fig 4. Kernel Density Estimation representation

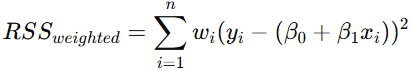
**KERNEL DENSITY BASED LINEAR REGRESSION:**

Core Idea:-

KDLR extends traditional regression by replacing equal weighting with density-based weighting. Data points in dense regions receive higher weights, while those in sparse or noisy regions are de-emphasized.

Mathematical Formulation:-

The weighted least squares objective for KDLR is:



Where the weights wi are derived from the KDE of the dataset:

wi ∝ f (xi)

**METHODOLOGY**

1. Data Pre-processing: Clean and normalize the dataset.
2. Kernel Density Estimation: Calculate the density f (xi) for each data point.
3. Weight Assignment: Normalize densities to obtain weights wi​.
4. Weighted Regression: Fit the regression model using weighted least squares.

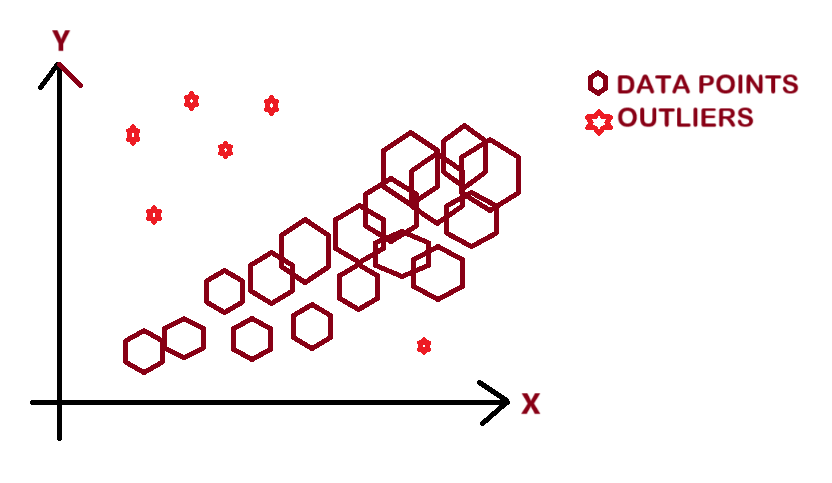


Fig 5. Increased the weights of denser points and decreased the weights of sparse points

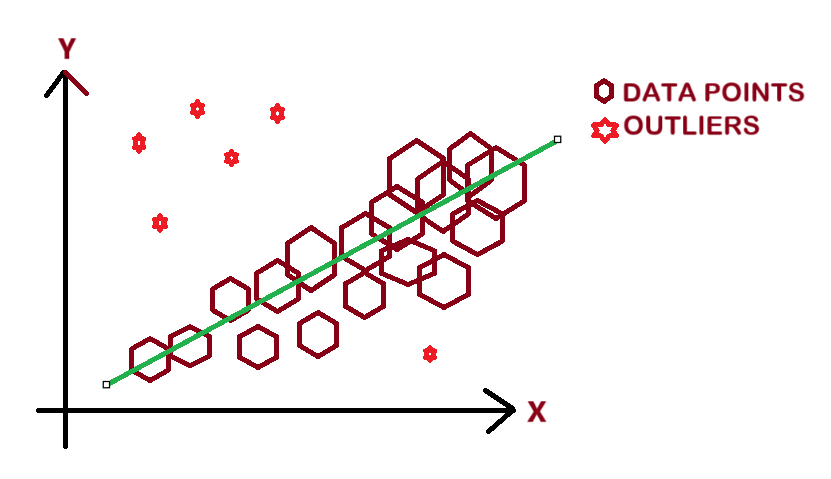


Fig 6. KLDR fits data with outliers

Kernels used:-

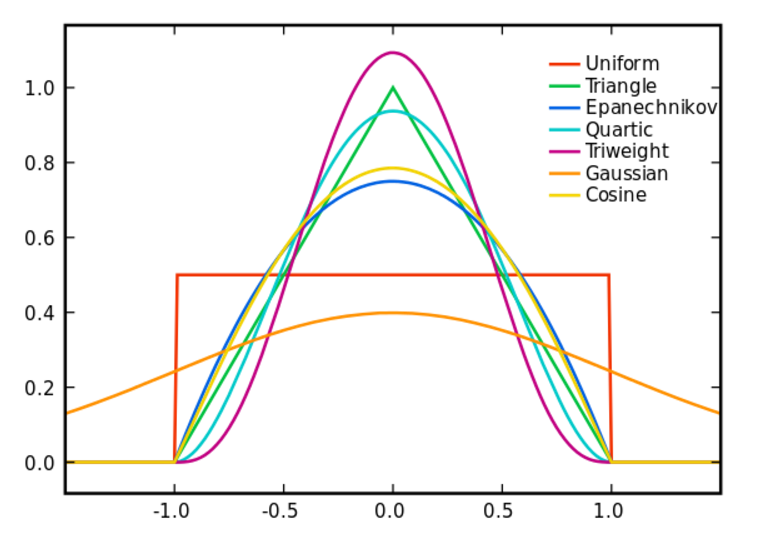


Fig 7. Kernel shapes

Gaussian Kernel:

A smooth, bell-shaped kernel that provides infinite support, making it ideal for continuous data, given by



Uniform Kernel:

A flat kernel that assigns equal weight to all points within its range, offering simplicity but less smoothness, given by



Triangular Kernel:

A linearly decreasing kernel that balances smoothness and computational efficiency, given by



Quartic Kernel:

A smooth kernel that emphasizes central points while providing finite support, given by



Tri-weight Kernel:

A highly smooth kernel that gives more weight to central points, ensuring high smoothness, given by



Cosine Kernel:

A kernel with a cosine shape that is smooth and finite in support, ideal for applications where moderate smoothness and compactness are desired, given by



**COMPARATIVE PERFORMANCE ANALYSIS**

Experiment Setup

1. Dataset: California housing dataset.
2. Scenarios: Introduced 50, 200, and 350 synthetic outliers.
3. Metrics: Evaluated mean squared error (MSE) and robustness to outliers.

**RESULTS**

* Linear Regression: Performance deteriorated significantly with increasing outliers.
* KDLR: Maintained low error and successfully identified underlying patterns.

Key Insights:-

* Outlier Resistance: Reduced sensitivity to sparse regions.
* Pattern Recognition: Effectively captured subtle, dense patterns in data.

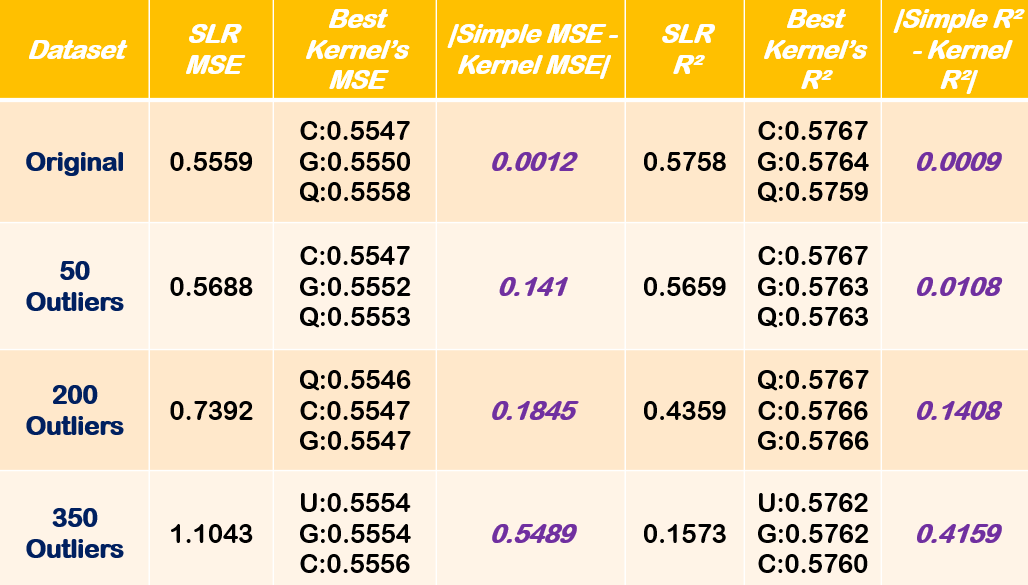


Fig 8. Results Comparison Table

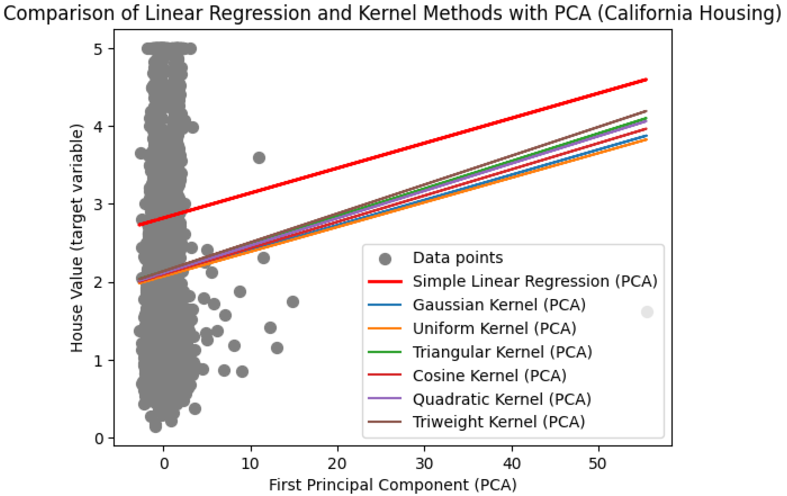


Fig 9. Multivariate Case Result (PCA)

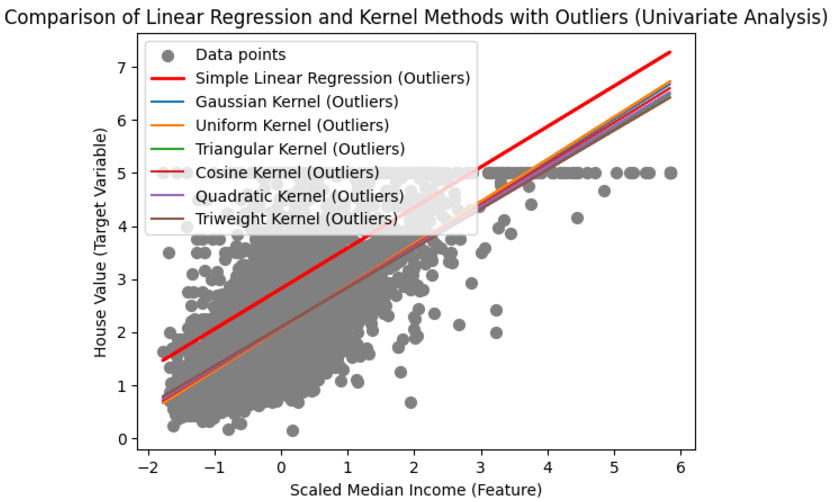


Fig 10. Univariate Case Result

**APPLICATIONS**

1. Finance: Robust stock price prediction in the presence of sparse anomalies.
2. Healthcare: Modelling patient data with rare conditions and outliers.
3. Environmental Science: Reliable climate modelling with noisy datasets.

KDLR’s adaptability extends to any domain where datasets exhibit heterogeneity or noise.

**CONCLUSION**

Key Takeaways:-

* Enhanced Robustness: KDLR reduces the influence of outliers, resulting in more reliable predictions.
* Adaptive Weighting: Effectively balances the contributions of dense and sparse regions.
* Pattern Recognition: Captures complex patterns often missed by traditional regression.

**FUTURE SCOPE**

* Hybrid Models: Incorporate KDLR into ensemble methods for improved generalization.
* Anomaly Detection: Extend KDLR to identify and interpret rare events in time-series data.
* Automation: Develop automated pipelines to integrate KDLR into real-world workflows.

**ACKNOWLEDGEMENT**

Our heartfelt gratitude to Professor Sung Hyuk Cha from Pace University for his guidance and support throughout this research on kernel density-based clustering. His invaluable insights and expertise have greatly enriched our understanding of advanced clustering techniques. We also extend our thanks to the Seidenberg School of Computer Science and Information Systems for providing the resources and encouragement to pursue this work. This study would not have been possible without the inspiration and mentorship of Professor Cha, whose dedication to fostering innovation has been instrumental in our learning journey.

**REFERENCES**

1. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
2. Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall.
3. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
4. Seber, G. A. F., & Lee, A. J. (2003). Linear Regression Analysis. Wiley.
5. Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric Regression. Cambridge University Press.
6. Fan, J., & Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapman & Hall.
7. Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (2000). Understanding Robust and Exploratory Data Analysis. Wiley.
8. Scott, D. W. (2015). Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley.
9. Huber, P. J. (1981). Robust Statistics. Wiley.
10. Koenker, R., & Hallock, K. F. (2001). "Quantile Regression". Journal of Economic Perspectives, 15(4), 143–156.
11. Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.
12. Chen, Y., & Liu, X. (2011). "Kernel Density Estimation". Encyclopedia of Database Systems. Springer.
13. Wasserman, L. (2006). All of Nonparametric Statistics. Springer.
14. Cleveland, W. S., & Devlin, S. J. (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting". Journal of the American Statistical Association, 83(403), 596–610.
15. Loader, C. (1999). Local Regression and Likelihood. Springer.