Kernel Density Estimated Linear Regression

Roshan N. Kalpavruksha[1], Rohan N. Kalpavruksha,[1], Teryn Cha,[2] Sung-Hyuk Cha,[1]
[1]Computer Science Department, Pace University, New York, NY, USA
[2]Computer Sciences, Essex County College, Newark, NJ, USA
rk68465n@pace.edu, rk25464n@pace.edu, yan@essex.edu, scha@pace.edu

Abstract

Regression analysis is a cornerstone of predictive modeling, with linear regression and kernel regression standing as two of its most prominent paradigms. However, each approach has inherent limitations: linear regression is highly susceptible to outliers in noisy and unevenly distributed datasets, while kernel regression often suffers from overfitting. When data exhibits linear trends, linear regression tends to generalize better, whereas kernel regression may fail to capture the broader patterns effectively. To address these challenges, we propose a novel methodology that integrates linear regression principles with kernel density estimation, termed Kernel Density Estimated Linear Regression (KDLR). This approach leverages kernel density estimation to assign higher weights to data points in dense regions, simultaneously deemphasizing the influence of sparse or noisy points. This dynamic weighting mechanism enhances robustness and improves the model's ability to recognize meaningful patterns. Extensive evaluations on datasets, including the California housing dataset with varying levels of outliers, demonstrate the efficacy of KDLR. Using the mean squared error metric, KDLR consistently achieves superior accuracy compared to traditional regression methods. By prioritizing dense data regions, it captures significant trends while mitigating the effects of noise. Applications of KDLR span diverse fields, including stock price prediction in finance, patient data modeling in healthcare, and climate modeling in environmental studies - domains where robust anomaly and noise handling are critical. This research establishes KDLR as a transformative tool for predictive analytics, offering a compelling blend of precision, resilience, and adaptability. Its ability to manage noisy data effectively while identifying critical patterns positions KDLR as a versatile and innovative solution for both academic research and industrial practice, with the potential to redefine best practices in regression modeling.

I. Introduction

Linear regression has been a cornerstone of statistical modeling and machine learning due to its simplicity and interpretability, as first conceptualized in (Galton 1886; Pearson 1896). Its fundamental assumption that all data points contribute equally to the regression model performs well for clean and uniformly distributed datasets. However, in real-world applications, datasets often deviate from this ideal, frequently containing noise and outliers. Such irregularities can significantly impact the performance of linear regression models. As illustrated in Figure 1 (a), even a single outlier can distort the fitted regression line, underscoring the vulnerability of ordinary linear regression to noisy data.

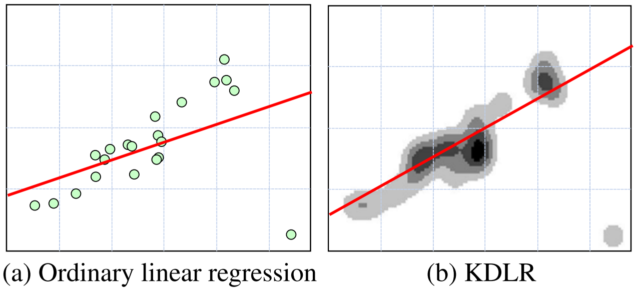
(a) Ordinary linear regression (b) KDLR

Figure 1: Motivation: Ordinary linear regression vs. KDLR.

Non-parametric density regression techniques, first introduced in (Nadaraya 1964; Watson 1964), address some of the limitations of linear regression by bypassing the need to explicitly estimate target functions or parameters during training. Instead, these methods rely on storing training instances and evaluating their relationships with a given query instance during prediction. While this approach offers flexibility, it is computationally intensive, resulting in slow predictions. Moreover, traditional kernel regression methods are prone to overfitting in sparse or noisy regions, particularly when kernel bandwidth is poorly calibrated. Their uniform treatment of all data points, irrespective of distribution density, further limits their robustness in heterogeneous datasets.

Weighted regression techniques, such as locally weighted scatterplot smoothing (LOWESS) (Cleveland 1979) and generalized additive models (GAMs) (Hastie and Tibshirani 1986), provide localized fitting by assigning weights based on proximity in the predictor space. While effective at improving local adaptivity, these methods fail to account for global data density in feature space. This omission often results in suboptimal performance when handling datasets with highly variable density distributions.

To address these challenges, this work proposes Kernel Density Estimated Linear Regression (KDLR), a novel methodology that integrates kernel density estimation into the linear regression framework. Unlike traditional methods, KDLR employs a dynamic, data-driven weighting mechanism that emphasizes dense regions while de-emphasizing sparse or noisy points, as depicted in Figure 1 (b). This approach enhances robustness against noise and ensures improved generalization compared to kernel regression, particularly when the data exhibits linear trends. Furthermore, by prioritizing dense regions, KDLR effectively captures meaningful patterns while mitigating the influence of outliers and noise - an underexplored capability in existing robust regression techniques.

Extensive evaluations demonstrate that KDLR consistently outperforms traditional regression methods in scenarios characterized by high noise and uneven data distributions. By combining the strengths of linear regression and kernel-based methods, KDLR represents a transformative advancement in predictive modeling. Its unique blend of precision, adaptability, and resilience not only deepens theoretical understanding but also offers significant practical applications in domains that demand robust and reliable data analysis.

The remainder of this paper is organized as follows: Section II provides a concise review of kernel regression and linear regression, followed by the introduction of the proposed Kernel Density Linear Regression (KDLR) approach. Section III presents experimental results using the California Housing, Adult Income, and Energy Efficiency datasets, with intentionally added noise to evaluate the robustness of the proposed model. Finally, Section IV concludes the work and discusses future directions.

Methodology

Kernel Regression

Kernel regression is a non-parametric method that predicts the target value of a query instance q by computing the kernel-weighted average of neighboring reference points. Let $\mathcal{R}$ denote the reference set, and $t(r)$ represent the target value for $r \in \mathcal{R}$. The kernel regression predictor is expressed as:

$$\text{KR}(q, \mathcal{R}) = \frac{\sum_{r\in\mathcal{R}} K(u)t(r)}{\sum_{r\in\mathcal{R}} K(u)} \tag{1}$$

$$\text{where} \quad u = \frac{d(x,y)}{h} \tag{2}$$

h is the bandwidth and $d(r, q)$ represents the distance between r and q.

Kernel Density Estimation (KDE), introduced by Parzen (Parzen 1962), is a widely utilized non-parametric approach for estimating the probability density function (PDF) of a dataset. Various kernel functions have been proposed in the literature, as summarized in (Silverman 1986; Botev, Grotowski, and Kroese 2010). In this study, seven commonly used kernel functions are evaluated: uniform, triangular, Epanechnikov, Triweight, cosine, exponential, and Gaussian kernels. For a normalized distance u, the kernel functions are defined as:

$$K_{\text{uni}}(u) = \begin{cases} \frac{1}{2} & \text{if } |u| \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$K_{\text{tri}}(u) = \begin{cases} 1-|u| & \text{if } |u| \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$K_{\text{epa}}(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{if } |u| \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$K_{\text{tw}}(u) = \begin{cases} \frac{35}{32}(1-u^2)^3 & \text{if } |u| \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$K_{\text{cos}}(u) = \begin{cases} \frac{\pi}{4}\cos\left(\frac{\pi u}{2}\right) & \text{if } |u| \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$K_{\text{exp}}(u) = \frac{1}{2}e^{-|u|} \tag{8}$$

$$K_{\text{gau}}(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} \tag{9}$$

Univariate Ordinary Linear Regression

Ordinary linear regression is a parametric approach to model the relationship between a single predictor x and a target variable T. The goal is to determine a line that minimizes the prediction error, defined as:

$$P(x, W) = w_1 x + w_0 \qquad \text{where } W = (w_0, w_1) \tag{10}$$

where w_1 and w_0 denote the slope and intercept, respectively.

The slope w_1 is determined by minimizing the mean squared error and can be expressed as:

$$w_1 = \frac{\mathbb{E}[XT] - \mathbb{E}[X]\mathbb{E}[T]}{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}. \tag{11}$$

The intercept w_0 is then computed using:

$$w_0 = \mu(T) - w_1\mu(X) \tag{12}$$

where $\mu(T)$ and $\mu(X)$ are the means of T and X, respectively.

Kernel Density Linear Regression

Kernel Density Linear Regression (KDLR) enhances ordinary linear regression by incorporating kernel density weighting. Let n be the number of samples in $\mathcal{R}$, i.e., $n = |\mathcal{R}|$. The slope w_1 in KDLR is computed as:

$$w_1 = \frac{\sum\limits_{x\in\mathcal{R}} xw(x)t(x) - n\sum\limits_{x\in\mathcal{R}} xw(x)\sum\limits_{x\in\mathcal{R}} w(x)t(x)}{\sum\limits_{x\in\mathcal{R}} x^2w(x) - n\left(\sum\limits_{x\in\mathcal{R}} xw(x)\right)^2}. \tag{13}$$

where

$$w(x) = \sum_{y\in\mathcal{R}} K\left(\frac{d(x,y)}{h}\right). \tag{14}$$

Let $\mu_k(X)$ and $\mu_k(T)$ be the kernel-weighted means of X and T:

$$\mu_k(X) = \frac{\sum\limits_{x\in\mathcal{R}} xw(x)}{n}, \tag{15}$$

$$\mu_k(T) = \frac{\sum\limits_{x\in\mathcal{R}} w(x)t(x)}{n}. \tag{16}$$

Using these kernel-weighted means, the intercept w_0 is given by:

$$w_0 = \mu_k(T) - w_1\mu_k(X). \tag{17}$$

Multivariate Kernel Density Linear Regression

The extension of KDLR to the multivariate case considers multiple predictors $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ to model the target variable T. The predictive function is now expressed as:

$$P(\mathbf{X}, W) = w_0 + \sum_{j=1}^{d} w_j X_j, \tag{18}$$

where $W = (w_0, w_1, \ldots, w_d)$ represents the regression coefficients.

The optimal coefficients w_j are computed by minimizing the kernel-weighted mean squared error, leading to the following expressions:

$$W = (X^T W_k X)^{-1} X^T W_k T, \tag{19}$$

where:

- X is the $n \times (d+1)$ design matrix including a column of ones for the intercept,
- T is the $n \times 1$ target vector,
- W_k is an $n \times n$ diagonal matrix with kernel weights $w(x_i)$ along the diagonal, defined as:

$$w(x_i) = \sum_{y\in\mathcal{R}} K\left(\frac{d(\mathbf{x}_i, \mathbf{y})}{h}\right). \tag{20}$$

The kernel-weighted means for multivariate features are computed as:

$$\mu_k(\mathbf{X}) = \frac{\sum\limits_{\mathbf{x}\in\mathcal{R}} w(\mathbf{x})\mathbf{x}}{n}, \tag{21}$$

$$\mu_k(T) = \frac{\sum\limits_{\mathbf{x}\in\mathcal{R}} w(\mathbf{x})t(\mathbf{x})}{n}. \tag{22}$$

Finally, the intercept term in the multivariate case is given by:

$$w_0 = \mu_k(T) - \sum_{j=1}^{d} w_j \mu_k(X_j). \tag{23}$$

This extension allows KDLR to effectively model multivariate relationships while maintaining robustness against outliers and irregular data distributions.

Mean Squared Error

Mean Squared Error (MSE) is a widely used metric for evaluating the performance of regression models by measuring the average squared difference between the actual and predicted values. It quantifies the extent to which the predicted values deviate from the true values, providing an indication of the model's accuracy. A lower MSE value signifies better predictive performance, as it indicates that the predictions are closer to the true values. Formally, given a dataset with n observations, where $t_i \in T$ represents the true value and $p_i \in P$ denotes the predicted value for the i-th instance, the MSE is defined as:

$$\text{MSE}(T, P) = \frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2. \tag{24}$$

This formulation ensures that larger errors contribute more significantly to the final error measure due to the squared term, making MSE particularly sensitive to outliers. MSE is commonly used in various regression tasks to compare models and optimize predictive accuracy.

Experiments

To evaluate the performance and robustness of the proposed models, we conducted experiments on three datasets: the California Housing Prices, Adult Income, and Energy Efficiency datasets. Each dataset was chosen for its unique characteristics, allowing for comprehensive analysis under varying conditions.

California Housing Prices

The California Housing Prices dataset, derived from the 1990 California census (Nugent 2017), comprises attributes relevant to predicting housing prices. For visualization purposes, Figure 2 illustrates the predictive surfaces generated

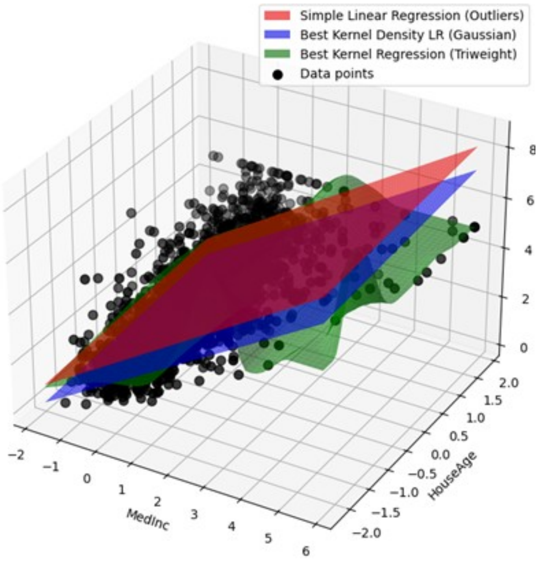

Figure 2: California Housing Prices: OLR, KR. and KDLR.

Table 1: Mean Squared Error (MSE) on California Housing Prices data with varying levels of outliers.

model \ outliers		Original	50	200	350
OLR		0.5559	0.5833	0.7321	1.1041
KR	Uniform	0.5125 (h=1.0)	0.7773 (h=2.0)	0.8739 (h=1.0)	1.3056 (h=1.0)
	Triangular	0.4800 (h=1.0)	0.7121 (h=2.0)	0.8719 (h=2.0)	1.2382 (h=2.0)
	Epanech-	0.4682 (h=1.0)	0.6984 (h=2.0)	0.8457 (h=2.0)	1.2098 (h=2.0)
	Triweight	0.4574 (h=1.0)	0.6909 (h=2.0)	0.8235 (h=2.0)	1.1906 (h=2.0)
	Cosine	0.4824 (h=1.0)	0.7165 (h=2.0)	0.8777 (h=2.0)	1.2461 (h=2.0)
	Exponen-	0.4346 (h=0.1)	0.5633 (h=0.1)	0.9228 (h=0.5)	1.2877 (h=0.5)
	Gaussian	0.5158 (h=0.1)	0.6919 (h=0.5)	0.8071 (h=0.5)	1.2173 (h=0.5)
KDLR	Uniform	0.5559 (h=5.0)	0.5562 (h=5.0)	0.5557 (h=5.0)	0.5557 (h=5.0)
	Triangular	0.5561 (h=5.0)	0.5558 (h=5.0)	0.5549 (h=5.0)	0.556 (h=5.0)
	Epanech-	0.5558 (h=5.0)	0.5555 (h=5.0)	0.5547 (h=5.0)	0.5559 (h=5.0)
	Triweight	0.5580 (h=5.0)	0.5570 (h=5.0)	0.5549 (h=5.0)	0.5561 (h=5.0)
	Cosine	0.5547 (h=5.0)	0.5549 (h=5.0)	0.5548 (h=5.0)	0.5558 (h=5.0)
	Exponen-	0.5551 (h=5.0)	0.5549 (h=5.0)	0.5547 (h=5.0)	0.5558 (h=5.0)
	Gaussian	0.5555 (h=5.0)	0.5554 (h=2.0)	0.5549 (h=5.0)	0.5556 (h=5.0)

by different models, using the `housing_median_age` and `median_income` attributes as input features.

Figure 2 highlights the predictive surfaces produced by Ordinary Linear Regression (OLR), Kernel Regression (KR), and Kernel Density Linear Regression (KDLR). While OLR and KDLR yield linear planes, KR demonstrates a more flexible, non-linear prediction surface, capturing complex relationships in the data.

The dataset consists of 20,640 samples, with 80% allocated for training and the remaining 20% for testing. The target variable is Median House Value, predicted based on eight features: MedInc (Median Income), HouseAge, AveRooms (Average Rooms per Household), AveBedrms (Average Bedrooms per Household), Population, AveOccup (Average Occupants per Household), Latitude, and Longitude.

Table 1 presents the Mean Squared Error (MSE) of different models under varying levels of artificially introduced outliers. KR exhibited excellent performance on the original dataset but suffered a significant decline in predictive accuracy as the number of outliers increased. Similarly, OLR's performance deteriorated sharply with higher levels of noise.

In contrast, Kernel Density Linear Regression (KDLR) demonstrated remarkable robustness to outliers. Across different kernel variations, particularly Gaussian and Cosine kernels, KDLR exhibited stable MSE values with minimal degradation, even in the presence of a high number of outliers. This resilience highlights KDLR's superior adaptability to noisy data.

As the level of contamination increased, the performance gap between OLR and KDLR widened significantly, underscoring KDLR's robustness. Unlike OLR and KR, which suffered considerable performance deterioration, KDLR maintained consistent accuracy, reinforcing its suitability for real-world scenarios where datasets often contain outliers. These results suggest that KDLR is a reliable alternative to traditional regression models, particularly when handling datasets susceptible to noise and outliers.

Adult Income

The Adult Income dataset (Becker and Kohavi 1996) is used to predict whether an individual's annual income exceeds a specified threshold. From the original 32,561 samples, a subset of 2,605 samples (8%) is selected for the training set, while 651 samples (2%) are allocated for testing. The target variable, `income_>50K`, represents whether an individual earns more than $50,000$ per year.

Table 2: Mean Squared Error (MSE) on Adult Income data with varying levels of outliers.

model		h	Original	20 outlie-	50 outlie-
OLR			0.1287	0.1426	0.1718
KR	Uniform	3	0.1582	0.2473	0.2819
		4	0.1569	0.1827	0.2693
	Triangular	3	0.1603	0.2378	0.3185
		4	0.1568	0.2060	0.2606
	Epanech-	3	0.1584	0.2374	0.3362
		4	0.1636	0.2214	0.2747
	Triweight	3	0.1611	0.2359	0.3656
		4	0.1665	0.2302	0.3037
	Cosine	3	0.1564	0.239	0.3096
		4	0.1598	0.2046	0.2559
	Exponen-	3	0.1725	0.175	0.1924
		4	0.1771	0.1798	0.1970
	Gaussian	3	0.1509	0.1553	0.1825
		4	0.1663	0.1698	0.1976
KDLR	Uniform	3	0.1287	0.1288	0.1288
		4	0.1287	0.1288	0.1288
	Triangular	3	0.1290	0.1290	0.1290
		4	0.1288	0.1288	0.1288
	Epanech-	3	0.1287	0.1287	0.1288
		4	0.1286	0.1287	0.1287
	Triweight	3	0.1289	0.1289	0.1289
		4	0.1287	0.1287	0.1288
	Cosine	3	0.1286	0.1287	0.1287
		4	0.1287	0.1287	0.1288
	Exponen-	3	0.1288	0.1293	0.1292
		4	0.1287	0.1296	0.1300
	Gaussian	3	0.1287	0.1292	0.1294
		4	0.1287	0.1301	0.1314

The dataset comprises 14 features, including age, work class, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country. These attributes collectively capture key demographic and occupational characteristics relevant to income prediction.

Table 2 presents the performance of Ordinary Linear Regression (OLR), Kernel Regression (KR), and Kernel Density Linear Regression (KDLR) across seven different kernel functions. KDLR achieves the lowest Mean Squared Error (MSE), demonstrating slightly improved predictive accuracy on the dataset without added outliers. However, OLR's MSE deteriorates rapidly as the number of outliers increases, highlighting its sensitivity to data contamination.

Among KDLR models, the Epanechnikov kernel with a bandwidth of $(h = 4)$ and the cosine kernel with a bandwidth of $(h = 3)$ achieve optimal performance, maintaining stability both with and without outliers.

In contrast, KR consistently underperforms relative to OLR, which may be attributed to suboptimal bandwidth selection. This observation underscores the sensitivity of kernel-based models to hyperparameter tuning, particularly the choice of bandwidth, which significantly influences predictive performance.

Energy Efficiency

The Energy Efficiency dataset (Tsanas and Xifara 2012) is used to assess the heating and cooling loads of buildings based on various architectural and environmental factors. The dataset consists of 768 samples, with 80% allocated for training and the remaining 20% for testing. The target variable is Heating_Load, predicted based on eight features: Relative_Compactness, Surface_Area, Wall_Area, Roof_Area, Overall_Height, Orientation, Glazing_Area, and Glazing_Area_Distribution.

Notably, Ordinary Linear Regression (OLR) outperforms Kernel Regression (KR) on this dataset, suggesting that the data exhibits a strong linear structure. This observation highlights the advantage of linear models when the underlying relationships align closely with linear assumptions. However, Kernel Density Linear Regression (KDLR), with an appropriately chosen kernel and bandwidth, surpasses both OLR and KR, demonstrating its flexibility in capturing underlying patterns in the data.

Table 3 presents the Mean Squared Error (MSE) results for OLR, KR, and KDLR across seven different kernel functions under varying levels of artificially introduced outliers.

Among the models evaluated, KDLR with a uniform kernel $(h = 5)$ achieves the lowest MSE, demonstrating superior predictive accuracy on the dataset without outliers. In contrast, OLR's performance deteriorates significantly as the number of outliers increases, highlighting its susceptibility to data contamination. While KDLR also experiences some performance degradation, its decline remains substantially less pronounced than that of OLR, reinforcing its robustness in handling noisy data.

Kernel Regression (KR) also exhibits resilience to outliers because the extreme values introduced into the dataset have minimal impact on the testing data. However, since KR does not assume linearity, its overall MSE remains higher compared to both KDLR and even OLR. This suggests that KR may be less suited for datasets where linear models already provide strong predictive power.

Table 3: Mean Squared Error (MSE) on Energy Efficiency data with varying levels of outliers.

model		h	Original	20 outli-	30 outli-
OLR			9.1532	12.241	15.379
KR	Uniform	2.5	10.41	10.41	10.41
		3	12.587	12.587	12.586
	Triangular	3	10.315	10.315	10.315
		3.5	11.970	11.970	11.970
	Epanech-	2.5	10.870	10.870	10.870
		3	9.664	9.664	9.663
	Triweight	3.5	9.9670	9.9670	9.9670
		4	11.283	11.283	11.283
	Cosine	3	10.529	10.529	10.529
		3.5	12.270	12.270	12.270
	Exponen-	0.5	9.075	9.075	9.075
		1	20.892	20.892	20.892
	Gaussian	1.5	15.817	15.817	15.817
		2	29.040	29.040	29.040
KDLR	Uniform	5	9.1203	9.4040	9.4409
		10	9.1532	9.5586	9.7470
	Triangular	5	9.1391	9.4395	9.4674
		10	9.1329	9.4875	9.5591
	Epanech-	5	9.1674	9.4468	9.4753
		10	9.1300	9.4667	9.5262
	Triweight	5	9.2123	9.4862	9.5232
		10	9.1239	9.4468	9.4954
	Cosine	5	9.1309	9.4166	9.4427
		10	9.1374	9.4950	9.5725
	Exponen-	2.5	9.1272	9.4517	9.4996
		3	9.1266	9.4575	9.5091
	Gaussian	2.5	9.1254	9.4075	9.4543
		3	9.1209	9.4124	9.4613

These findings underscore the robustness and adaptability of KDLR in modeling energy efficiency data, particularly in the presence of non-linear dependencies. The integration of kernel density estimation into linear regression proves highly effective for capturing complex patterns and achieving precise, reliable predictions in energy efficiency modeling.

Conclusion

In this study, we introduced Kernel Density Linear Regression (KDLR), a novel hybrid approach that integrates kernel density estimation with linear regression to address the limitations of traditional regression methods. The proposed KDLR dynamically assigns data-driven weights based on density, enabling it to emphasize dense regions while deemphasizing sparse or noisy areas. This mechanism enhances its robustness to outliers and noise while maintaining the interpretability and simplicity of linear regression.

Through extensive experiments on three diverse datasets

- California Housing, Adult Income, and Energy Efficiency - we demonstrated the superior performance of KDLR. The results consistently showed that KDLR outperformed both Ordinary Linear Regression (OLR) and Kernel Regression (KR) in terms of predictive accuracy and robustness, particularly under challenging scenarios involving outliers and uneven data distributions.

Key observations include the following:

- In the California Housing dataset, KDLR maintained stable performance even as the number of outliers increased, outperforming both OLR and KR in terms of Mean Squared Error (MSE).
- For the Adult Income dataset, KDLR achieved slight but consistent improvements over OLR and KR across various kernel functions, showcasing its adaptability to real-world data characteristics.
- In the Energy Efficiency dataset, KDLR demonstrated its ability to model complex relationships with exceptional precision, achieving the lowest MSE among all methods.

Overall, KDLR establishes itself as a robust and effective regression technique, bridging the gap between linear and kernel-based methods. Its capacity to adapt to heterogeneous data distributions and mitigate the influence of outliers makes it a valuable tool for predictive modeling in diverse domains.

Future work will explore the extension of KDLR to multivariate settings and its application to larger-scale datasets. Additionally, optimizing kernel bandwidth selection and exploring alternative weighting mechanisms can further enhance its performance and applicability.

References

Becker, B., and Kohavi, R. 1996. Uci machine learning repository: Adult [dataset]. Accessed: 2025-01-27 `https://doi.org/10.24432/C5XW20`.

Botev, Z. I.; Grotowski, J. F.; and Kroese, D. P. 2010. Kernel density estimation via diffusion. *The Annals of Statistics* 38(5):2916–2957.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74(368):829–836.

Galton, F. 1886. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 15:246–263.

Hastie, T., and Tibshirani, R. 1986. Generalized additive models. *Statistical Science* 1(3):297–310.

Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability and Its Applications* 9(1):141–142.

Nugent, C. 2017. California housing prices dataset. Accessed: 2025-01-27 `https://www.kaggle.com/datasets/camnugent/california-housing-prices`.

Parzen, E. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* 33(3):1065–1076.

Pearson, K. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. A* 187:253–318.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Tsanas, A., and Xifara, A. 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings* 49:560–567. `https://api.semanticscholar.org/CorpusID:109658267`.

Watson, G. S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 26(4):359–372.