Atefe Rajabi – 40230563

Problem Definition

Dataset Visualization

Preprocessing

Derivation-Norm2

SkLearn Library

Simulated Annealing

Alternate Search

Discussion

Problem Definition

The chosen dataset for this project is the Boston Housing Dataset, which involves predicting the median value of owner-occupied homes (MEDV) in $1000s based on selected housing-related features. This project aims to explore a regression problem where the objective is to develop a model that accurately estimates housing prices using a subset of features.

Features Selected:

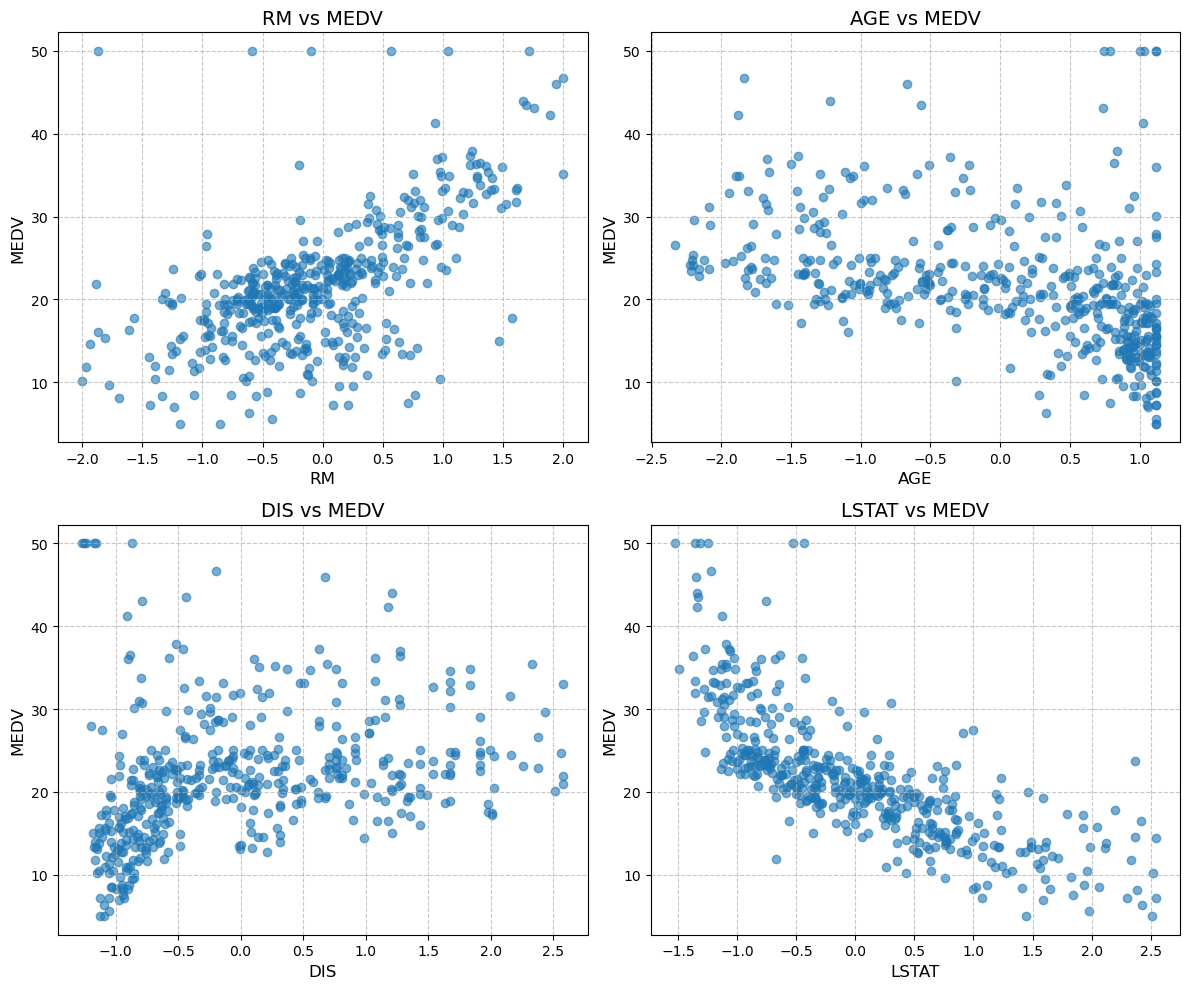
1. RM: Average number of rooms per dwelling.
2. AGE: Proportion of owner-occupied units built before 1940.
3. DIS: Weighted distances to five Boston employment centers.
4. LSTAT: Percentage of the population considered lower status.

Problem Definition:

Using the selected features (RM, AGE, DIS, and LSTAT), the goal is to:

1. Build a regression model that predicts the target variable MEDV (Median value of owner-occupied homes in $1000s).
2. Determine the optimal coefficients and offsets for the model based on the training set (80% of the dataset).
3. Analyze the contribution and impact of each feature separately by assessing the individual errors associated with them.

Dataset Visualization

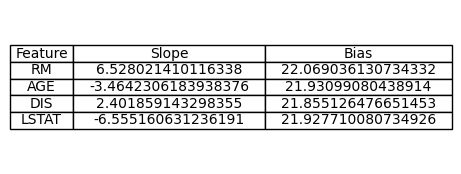
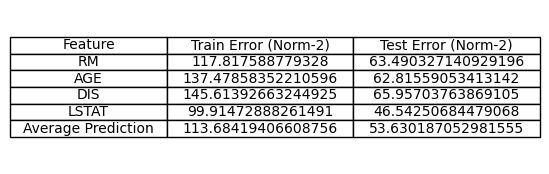


preprocessing

The preprocessing of the dataset involves two key steps to prepare the data for analysis. First, the selected features are standardized to have a mean of zero and a standard deviation of one. This ensures that all features are on the same scale, which is particularly important for models sensitive to the magnitude of the data. Second, an outlier removal process is applied using the interquartile range (IQR) method. This method identifies and removes data points that fall outside 1.5 times the IQR from the first and third quartiles for each feature, reducing the influence of extreme values.

Linear regression

To apply linear regression, it is necessary to minimize an error function, which can be based on different norms such as L0, L1, L2, or L∞. For the L2 norm, the error function is differentiable, allowing us to derive it with respect to the model parameters. Using gradient descent, we can iteratively optimize the parameters to find the best fit. However, for L0, L1, and L∞ norms, the error functions are not differentiable, requiring alternative optimization approaches. In this context, methods such as simulated annealing or alternate search strategies are applied to optimize the model parameters effectively.

Derivation-Norm2



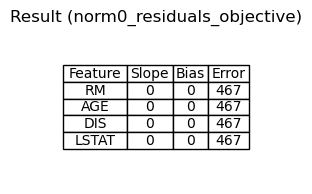
SKLearn Library

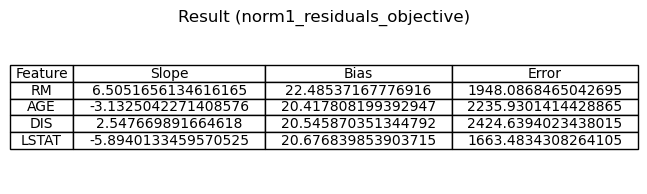
Applying linear regression on whole features once.

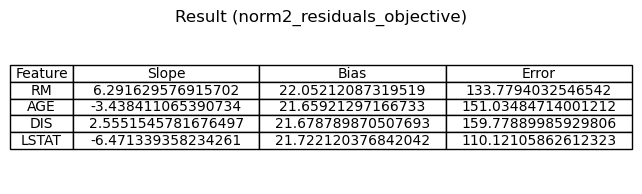
Mean Squared Error: 46.174958773873314 on test data

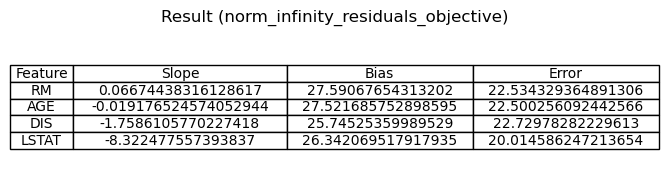
Coefficients: [ 2.94198891 -0.48296217 -0.96901766 -5.27106164]

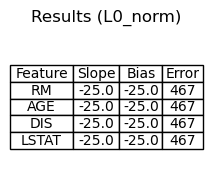
Intercept: 22.02022636917157

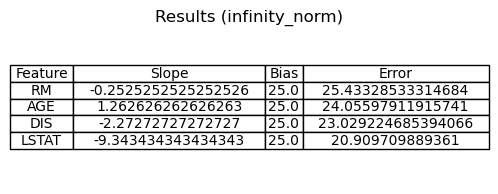
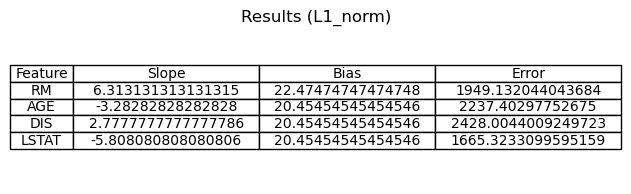
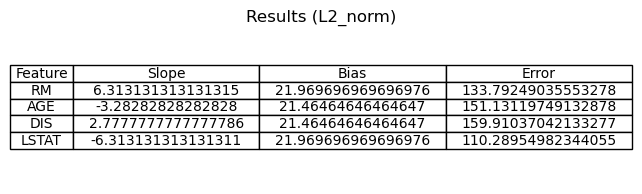
Simulated Annealing







Alternate Search

Discussion

Considering features separately and then averaging the predictions is less effective than applying linear regression with all features at once because it disregards the relationships and interactions among the features.

1. **Lack of Feature Interaction**:  
   Linear regression optimizes all coefficients simultaneously, capturing the combined effect of multiple features on the target variable. When features are considered independently, their interactions are ignored, leading to suboptimal predictions.
2. **Bias in Coefficients**:  
   By training models on individual features separately, the coefficients are determined without considering the contributions of other features. This can result in biased estimates since real-world relationships between features and the target variable are often interconnected.
3. **Averaging Dilutes Accuracy**:  
   When predictions from individual features are averaged, the model essentially assumes equal importance for all features, which is not always the case. In contrast, linear regression assigns weights to each feature based on their contribution to minimizing the overall error.
4. **Redundancy**:  
   Some features may provide overlapping or redundant information. Training on all features together allows the model to appropriately allocate weights, reducing overemphasis on any one feature.
5. **Minimization of Error**:  
   Linear regression works by minimizing a single global error function (e.g., norm-2 or MSE) across all features simultaneously. When features are treated separately, errors are minimized independently, but the combined predictions may not achieve the same level of overall optimization.

In the provided results, the **train and test errors** for the averaged predictions are higher (53.63) compared to the test error of 46.17 achieved when all features were used together. This highlights that applying linear regression on the full set of features yields better accuracy because it takes into account their combined effects and optimizes the error function globally.

The error in the infinity norm is often less than the error in other norms, due to the way each norm measures error. This norm aims to minimize the worst-case error rather than the sum or average of errors. As a result, it ignores smaller deviations in favor of controlling the largest error. By concentrating on the maximum error, the norm can lead to lower values compared to other norms because it sacrifices performance on smaller errors to minimize the peak error.