Project Report

Project Title: Linear Regression on Boston Housing Dataset

Name : Ranjesh Kumar Roy

Department : Mathematics

Roll No : 2521MA01

Abstract

This project applies **Linear Regression** techniques to the Boston Housing dataset to analyze and predict the median value of owner-occupied homes (MEDV). The model identifies relationships between socio-economic and environmental predictors such as crime rate, number of rooms, pupil-teacher ratio, and tax rates. Results demonstrate that variables like the average number of rooms per dwelling (RM) and percentage of lower status population (LSTAT) are the most influential predictors. The project provides insights into regression modeling, evaluation, and interpretation of housing price determinants.

Introduction

Regression analysis is one of the fundamental techniques in machine learning and statistics. **Linear Regression**, in particular, is widely used for modeling relationships between dependent and independent variables. This project explores the use of Linear Regression on the Boston Housing dataset to predict housing prices and to analyze key factors influencing property values.

Linear Regression: A Brief Overview

Linear Regression assumes a linear relationship between the dependent variable (**Y**) and independent variables (**X**).

The general form of the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

- **Y** = dependent variable (target, here `MEDV`)
- $\mathbf{X_i}$ = independent predictor variables
- $\boldsymbol{\beta_i}$ = regression coefficients (parameters)
- $\boldsymbol{\beta_0}$ = intercept
- $\boldsymbol{\epsilon}$ = error term

The objective is to estimate coefficients β by minimizing the **Residual Sum of Squares (RSS)**:

$$RSS = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) \right)^2$$

The solution for β is obtained using the **Normal Equation**:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Methodology

The methodology adopted in this project involves the following steps:

1. **Dataset**: The Boston Housing dataset consists of 506 rows and 15 columns.
    - Target variable: `MEDV` (Median value of homes).
    - Predictors: CRIM (crime rate), RM (average rooms), LSTAT (lower status %), TAX (property tax), PTRATIO (pupil-teacher ratio), NOX (nitric oxides concentration), etc.

```
Shape: (506, 15)
Columns: ['Unnamed: 0', 'crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax', 'ptratio', 'black', 'lstat', 'medv']
```

	Unnamed: 0	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0	1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	3	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	5	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

```
Target: medv
Predictors: ['Unnamed: 0', 'crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax', 'ptratio', 'black', 'lstat']
```

2. **Data Preprocessing**:

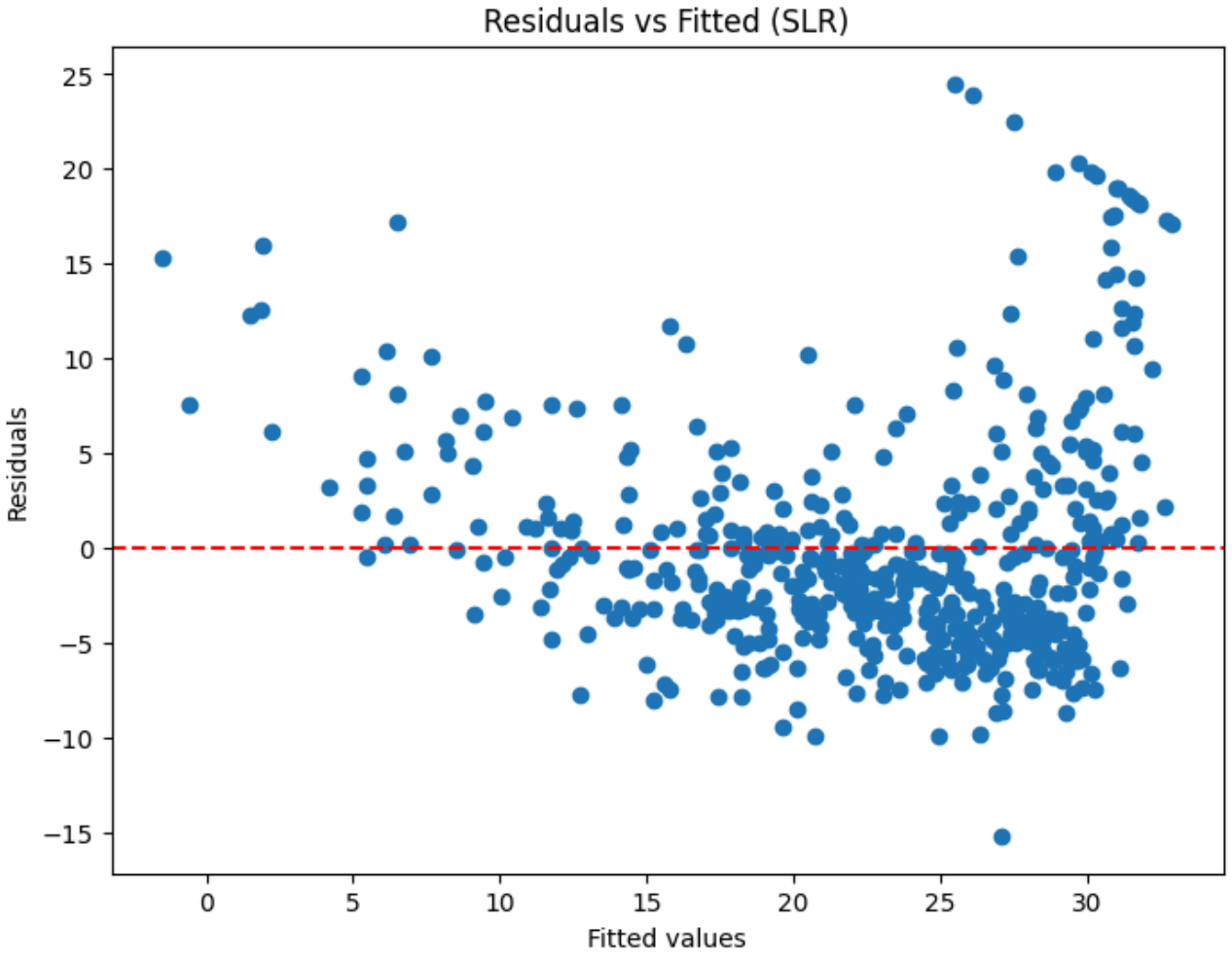
    - Dropped the index column.
    - Handled numerical variables.
    - Checked for correlations among predictors.
3. **Exploratory Data Analysis (EDA)**:
    - Distribution plots for MEDV.
    - Scatter plots of RM vs. MEDV and LSTAT vs. MEDV showed clear linear trends.
    - Correlation matrix confirmed RM (positive) and LSTAT (negative) as dominant features.
4. **Model Fitting**:
    - Implemented Multiple Linear Regression using **scikit-learn's** `LinearRegression`.
    - Estimated regression coefficients for each predictor.
5. **Evaluation**:
    - Used **R^2 score** to measure model performance.
    - Conducted residual analysis and QQ plots to validate assumptions of linear regression.

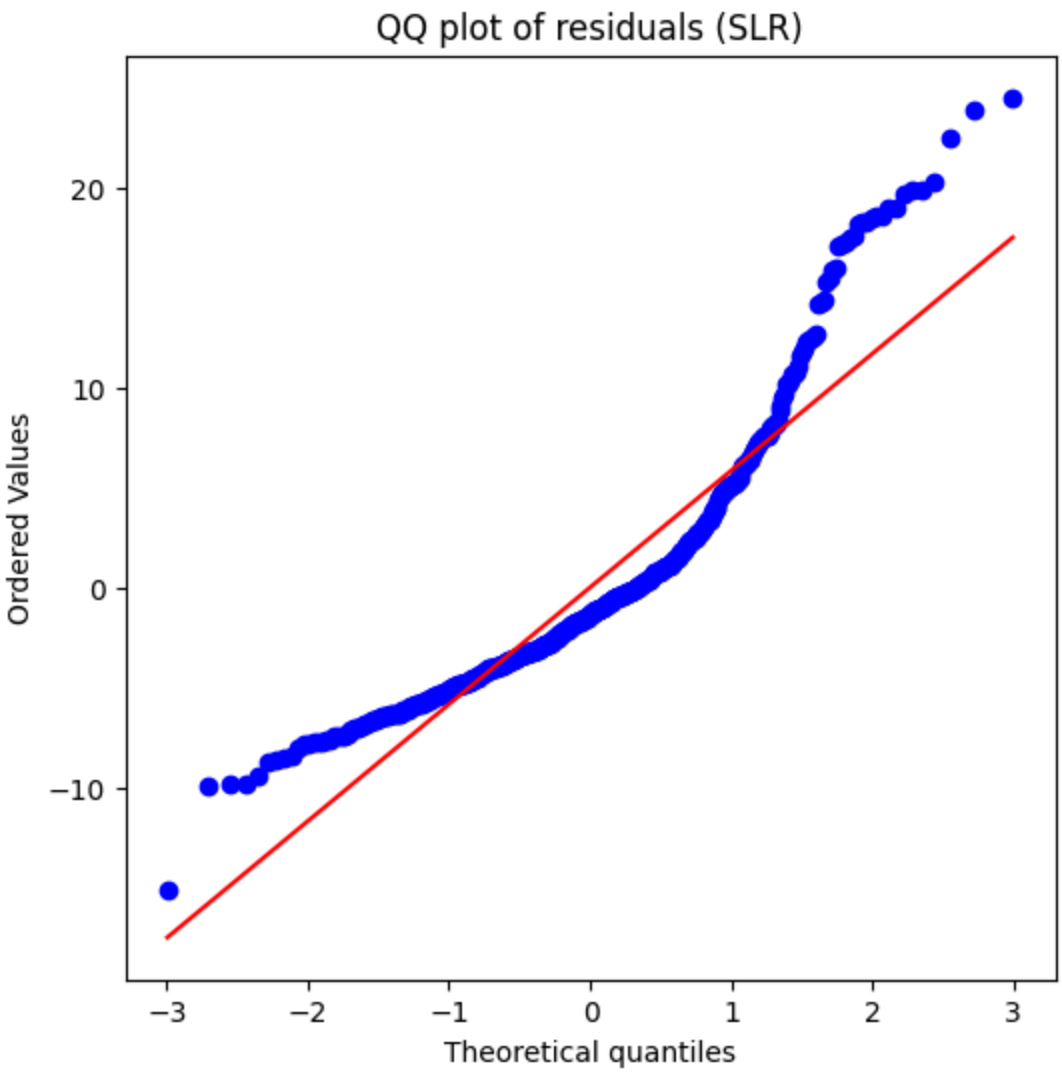
Results

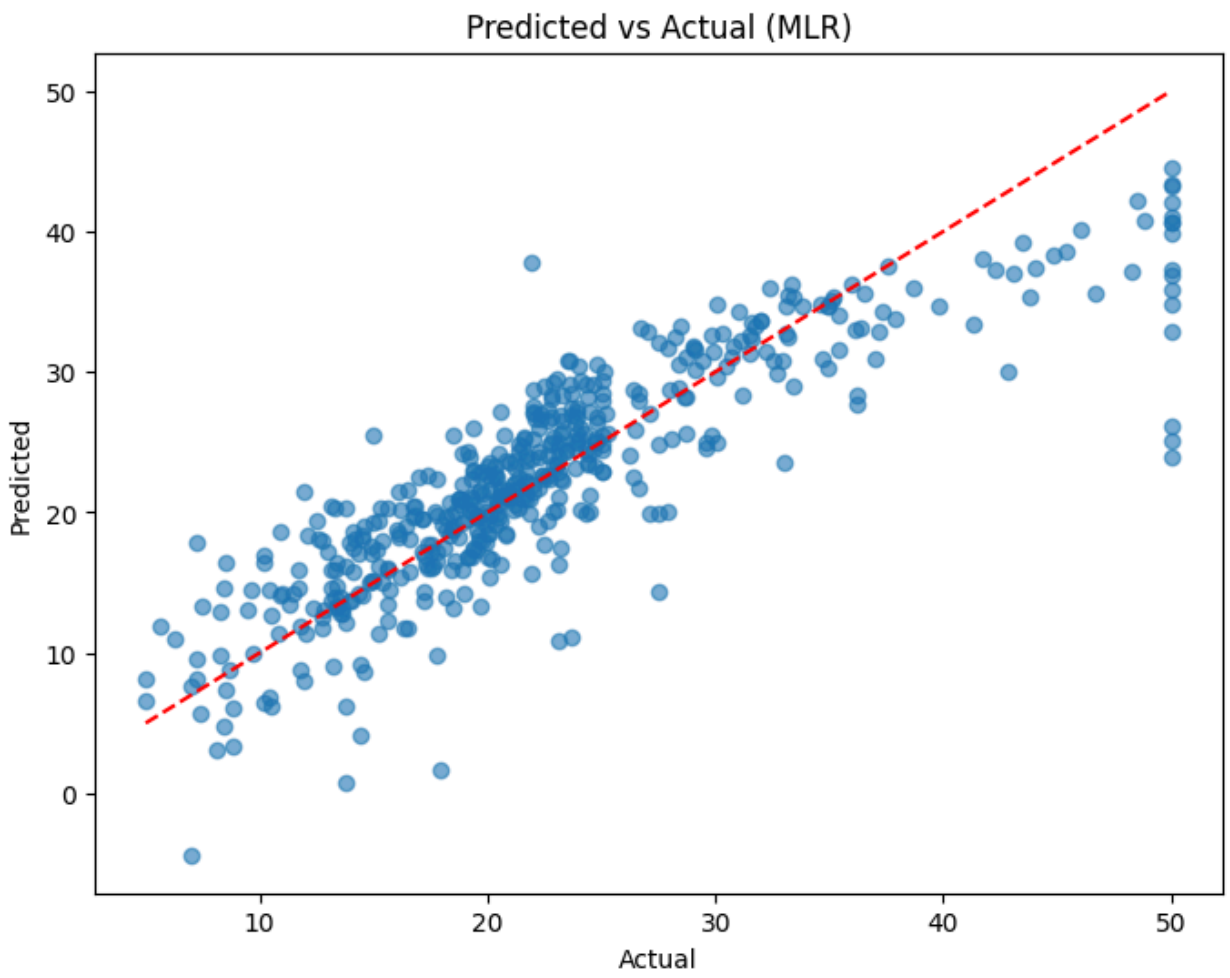
The regression model produced the following results:

- **Coefficient of Determination (R^2): ~0.74**
  → This means that ~74% of the variance in housing prices is explained by the predictors.

```
First 10 MLR coefficients: [ 3.64613519e+01 -2.52625878e-03 -1.08762336e-01  4.80307622e-02
  1.99323137e-02  2.70524534e+00 -1.75416021e+01  3.83922506e+00
 -1.93844593e-03 -1.49330389e+00]
MLR R^2 = 0.7414, Adjusted R^2 = 0.7340
```




QQ plot of residuals (SLR)
Ordered Values
Theoretical quantiles
20
10
0
−10
−3
−2
−1
0
1
2
3



- **Key Predictors:**
    - **RM (average rooms per dwelling):** Strong positive effect on MEDV.
    - **LSTAT (percentage of lower status population):** Strong negative effect on MEDV.
    - **NOX (air pollution) and TAX (property tax):** Moderate effects.
- **Residual Analysis:**

- Residuals were centered around zero, confirming unbiased predictions.
- Slight deviations from normality were observed in QQ plots, but overall assumptions were reasonably satisfied.

Conclusion

This project successfully demonstrated the use of Linear Regression to analyze the Boston Housing dataset. The results showed that socio-economic and environmental factors strongly influence housing prices. The model achieved good predictive performance with $R^2 \approx 0.74$.

Future work could extend the analysis using advanced regression methods or machine learning models to improve prediction accuracy.