Project 3: Fundamental of Statistics and Statistical Models

1 Research Question

The primary research question of this project is: Which factors significantly influence Boston home prices, and how can a linear regression model predict housing values?

2 Exploratory Analysis

A scatter plot (Figure 1) was used to explore the relationship between **rm** (average rooms per dwelling) and **medv** (housing prices). The plot reveals a positive linear relationship, with some outliers present. This supports the use of a linear regression model to capture the relationship between variables. The choice of a linear model is justified because the relationship between **rm** and **medv** is approximately linear, and the model provides interpretable results for assessing the influence of predictors on housing prices.

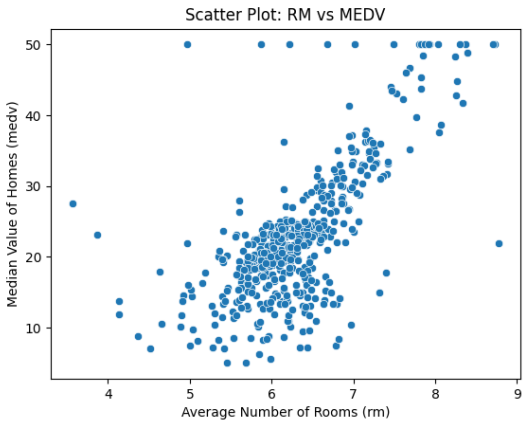

Figure 1: Scatter Plot of RM vs MEDV.

3 Data Splitting

The dataset was split into a training set (90%) and a test set (10%) using a random seed based on the last 4 digits of the student ID. This approach ensures the model is trained on one subset and tested on another, allowing for an unbiased evaluation of its performance on unseen data.

4 Linear Regression

A linear regression model was fitted using **rm** (average number of rooms per dwelling), **lstat** (percentage of lower socioeconomic population), and **crim** (per capita crime rate) as predictors, with **medv** (median home value) as the target variable. The training data (90% of the dataset) was used to estimate the model parameters via the ordinary least squares (OLS) method. The intercept ($\beta_0 = -3.87$) represents the predicted median home value when all predictors are zero. The coefficient for **rm** ($\beta_1 = 5.41$) indicates that for each additional room, the median home value increases by approximately \$5,415, holding other variables constant. Similarly, the coefficient for lstat ($\beta_2 = -0.52$) implies that a 1% increase in the lower socioeconomic population results in a \$520 decrease in the median home value. The coefficient for crim ($\beta_3 = -0.11$) suggests that higher crime rates have a slightly negative impact on housing prices.

5 Bootstrap

A bootstrap procedure with 1,000 iterations was performed to estimate confidence intervals for the model coefficients. Random samples with replacement were drawn from the training data, and an OLS regression model was fitted in each iteration. Confidence intervals were computed using the 2.5th and 97.5th percentiles of the bootstrap estimates. The confidence interval for **indus** $[-0.19, 0.06]$ includes 0, indicating it is not statistically significant at the $\alpha = 5\%$ level. Conversely, the confidence interval for **rm** $[3.66, 7.06]$ does not include 0, confirming a statistically significant positive relationship, where each additional room increases the median home value by approximately \$5,415. This analysis highlights the reliability of significant predictors ('rm', 'lstat', 'crim') and demonstrates the robustness of the model's coefficients.

6 Model Evaluation

The model's predictive performance was assessed using the Mean Squared Error (MSE) on the test data. The MSE was computed as:

$$\text{MSE} = \frac{1}{n}\sum(Y_{\text{test}} - Y_{\text{pred}})^2,$$

where Y_{test} represents the actual housing prices and Y_{pred} represents the predicted prices. The model achieved an MSE of 25.23, reflecting reasonable predictive accuracy on unseen data. This evaluation demonstrates the model's ability to generalize effectively and provide accurate predictions of housing prices.

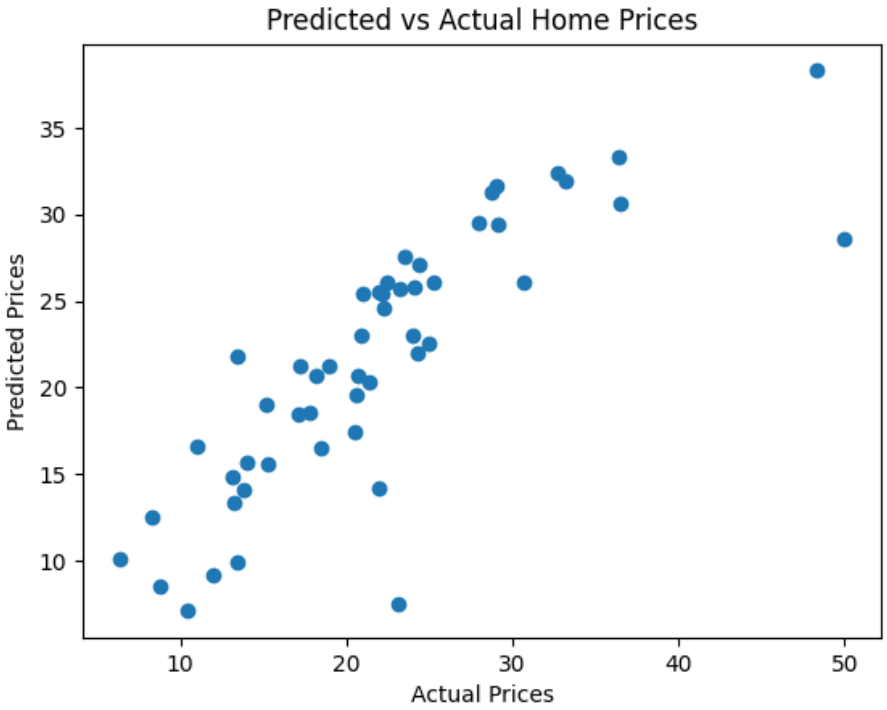

Figure 2: Predicted vs. Actual Home Prices. Most points are near the diagonal, demonstrating strong predictive performance.