Project 3: Fundamental of Statistics and Statistical Models

1 Research Question

The goal of this research is to build a predictive model for housing prices based on various features, such as the average number of rooms and the percentage of lower-status population. By fitting a linear regression model to the data, the report aims to understand the key factors influencing housing prices and evaluate the model's predictive performance.

2 Exploratory Analysis

To gain a clearer understanding of the dataset and the connections between the predictor variables and housing prices, i conducted an initial exploratory analysis. This analysis aims to visualize the distribution of important variables, spot potential outliers, and support the decision to use a linear regression model.

Visualization of variables

A scatter plot of rm (average number of rooms) vs. medv (housing prices) was created to explore the relationship between these variables:

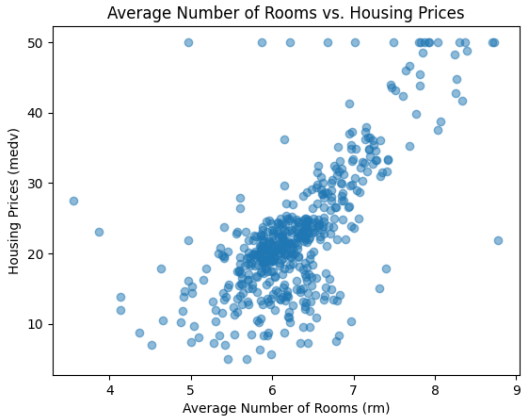

Figure 1: Scatter plot

This plot illustrates a positive linear relationship between the number of rooms and housing prices, indicating that houses with more rooms generally have higher prices. This visual evidence reinforces the appropriateness of using a linear regression model, as the relationship seems linear and can be effectively modeled with a straight-line equation..

Justification for linear model

The selection of a linear regression model is supported by the relationships observed in the dataset. The scatter plot shows a generally linear connection between predictors such as rm and the target variable medv, indicating that a linear model can effectively represent these trends. Additionally, linear regression offers a straightforward and interpretable way to measure the impact of predictors on housing prices. Given that the aim is to analyze and forecast housing prices based on these predictors, a linear model is well-suited for this purpose.

3 Data Splitting

Splitting the data into a training set and a test set enables the model to learn from one part of the data while being evaluated on another, unseen part. This method offers an impartial assessment of the model's

performance and its ability to generalize to new data.

4 Linear Regression

Introduction

A linear regression model was created using the training dataset to predict housing prices (medv) based on predictors like rm (average number of rooms) and lstat (percentage of lower-status population). The model was trained on 90% of the data, while the remaining 10% was set aside for evaluation..

Model fitting

The model utilized ordinary least squares (OLS) to reduce the residual sum of squares between the observed and predicted values of medv. This method helps to determine the coefficients that establish the linear relationship between the predictors and housing prices.

Coefficient interpretation

The intercept beta0 indicates the predicted housing price when all predictors are set to zero. Although this scenario may not be practical since variables like rm and lstat can't actually be zero—it serves as a baseline for the model. The coefficient for rm, beta5, measures the change in housing price for each additional room, assuming all other factors remain constant. For instance, if beta5 equals 4, it suggests that adding one room raises the median housing price by $4,000. In contrast, the coefficient for lstat is usually negative, signifying that a higher percentage of lower-status populations correlates with lower housing prices.

5 Bootstrap

The bootstrap method was employed to estimate the confidence intervals for the model coefficients. We resampled the training data with replacement 1,000 times, fitting the model to each resampled dataset and noting the coefficients. The 95% confidence intervals for each coefficient were then derived from the 2.5th and 97.5th percentiles of the bootstrap distribution.

Significance of beta5 (Coefficient for rm).

The coefficient for rm (average number of rooms) was determined to be statistically significant because its bootstrap confidence interval did not encompass zero. This suggests that the number of rooms positively influences housing prices, with each extra room raising the predicted price by the value of beta5.

6 Model Evaluation

To evaluate the performance of the linear regression model, we calculated the Mean Squared Error (MSE) with the test data. Assessing the model on unseen test data is essential, as it indicates how well the model can generalize to new data and offers an accurate measure of its predictive performance.
The MSE is calculated as the average of the squared differences between the true values Y_{test} and Y_{pred}.

$$\textbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_{\text{test}}[i] - Y_{\text{pred}}[i])^2$$

The model's performance is assessed using the mean squared error (MSE), which measures how close the predicted values are to the actual values. A lower MSE indicates better performance.
To evaluate the model, we used the test data to see how accurately it predicts housing prices for new, unseen instances. By calculating the MSE, we can objectively gauge the model's accuracy and determine its effectiveness for future predictions.