Analyzing the Complexities of Predicting the Median Value of Owner Occupied Homes in the 1970s Boston Housing Market Using Multiple Linear Regression Techniques

by

Michael Legard

Contents

Chapter 1

Introduction

The Boston housing market data set stands as a vital resource for extracting valuable insights into the complex world of real estate and urban development. Comprised of high-quality data that encompasses several important socioeconomic attributes and housing characteristics, this data set is highly relevant for various types of research. Through robust statistical modeling techniques, it offers the opportunity to uncover critical relationships and make informed decisions that can be applied to urban planning, property valuation, public policy, and housing market dynamics.

In the realm of data science and predictive analytics, the Boston housing market data set has been a cornerstone for testing and refining statistical and machine learning models. It holds crucial details about houses and neighborhoods in Boston, covering statistics like crime rates, land availability, and proximity to workplaces. As we dive into this study, the goal is not merely to analyze numbers but to derive a comprehensive understanding of the variables shaping house prices. By applying advanced regression techniques to this data set, this research is aimed at examining the implications of these insights to scrutinize and interpret their effects on broader housing markets.

This research endeavor is not just theoretical; the Boston housing market

data set provides real-world data, anchoring the resulting exploration as evidence of actual housing trends which offer concrete insights that resonate with the daily experiences of residents and potential home buyers in Boston. The aim is to not only understand the historical trends but to project forward, contributing to the creation of predictive models that can anticipate future shifts in housing markets but to identify the key variables that have continuously had significant impacts on housing prices with aspirations to equip anyone who may desire knowledge of these relationships with the tools needed to make informed decisions. Driven by this commitment to extract meaningful insights that transcend statistical abstraction and find practical applications, the goals of this research cast a broad net.

Primary objectives for this thesis revolve around the utilization of Multiple Linear Regression (MLR) techniques to understand the complex dynamics of the Boston Housing Market that are represented by the data set. The research begins with the optimization of a MLR model - a meticulous process of refinement aimed at enhancing the model's predictive accuracy which is later validated using an untrained subset of the original data. This goal of fine-tuning the model to adeptly capture the nuances within the data set will ultimately create a general tool for predicting house prices. The goal is not merely prediction; it's understanding the unique influence of each variable on house prices within the context of the model and drawing conclusions based on these influences. With these research goals in mind, a basic background of the terminology, mathematical principles, and notations are given to understand how these objectives are achieved at various stages of research.

Chapter 2

Background

In the vast ocean of statistical analysis methods, regression stands as a powerful and versatile method that can be employed to understand the relationships between variables. Regression encompasses a spectrum of techniques that can capture complex associations within data. At its essence, regression aims to model the connection between a dependent variable and one or more independent variables. This modeling allows us to predict or estimate the value of the dependent variable based on the known values of the independent variables. As the intricacies of regression analysis are explored in the context of the Boston housing market, a solid foundation in these statistical principles defining the specifics of the regression model is indispensable and dependent on a clear understanding of many fundamental concepts and terminology. This section serves to establish a foundation in the mathematical principles that are essential for comprehending the intricacies related to the type of statistical analysis performed in this research.

One notable concept that is a persistent issue with this sort of analysis is that it is often impractical or resource-intensive to gather exhaustive data over a vast population. When dealing with such populations or complex real-world scenarios, the collection of data is often an insurmountable challenge. Therefore, statisticians use a variety of techniques to estimate data of interest. This

is one of the key characteristics of regression and needs to be constantly taken into consideration to understand the concepts, notations, and implications of the regression fundamentals described below. To address this, the notation commonly used to describe an estimate is by $\hat{}$, which will be seen frequently for the duration of this research. With this kept in mind, some of the most basic relevant terminology to understanding the goals of this research are introduced.

Terminology and Mathematical Notation

- Linear regression: a statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables. The model assumes a linear relationship between the variables, aiming to find the best-fitting linear equation that describes the effect of the independent variables on the dependent variable.

- Multiple Linear Regression: an extension of linear regression, MLR models the linear relationship between a dependent variable and two or more predictors, allowing for a more nuanced understanding of how various factors collectively influence the response variable.

- Data set: an organized set or collection of data. Typically represented by a table, a data set contains information or records related to a specific topic of interest. A data set consists of rows which are used to identify particular instances of the data set, and columns which correspond to unique variables that can be of various data types.

- Observation: An observation refers to a single data point or measurement within a data set and represents a specific instance, event, or unit of

analysis (i.e. a particular row in a data set). One observation contains one value for each variable.

- Population: All individuals/cases that are of interest. In the context of this research, the population can be considered the entirety of the Boston housing market, for which thousands of observations are not included in the data set.

- Sample: a subset of the population used to collect data on. In the context of this research, the sample is the Boston housing market datastet. Samples are used to estimate effects that can then be applied to populations.

- Distribution: A distribution refers to the pattern or arrangement of values in a data set, indicating how frequently different values occur. It describes the spread of data and can take various shapes.

- Dependent variable: a variable whose value is dependent on the values of other variables. Also called the response variable, it is what research attempts to measure, understand, or predict in an experiment or analysis.

- Independent variable: Also known as a predictor, this type of variable is a factor that is purposefully recorded to observe its impact on the dependent variable. Also known as the explanatory variable.

- Quantitative variable: also called a 'numeric' variable, a quantitative variable is a variable whose data is measured numerically and generally over a continuous scale.

- Qualitative variable: also called a 'categorical' variable, a qualitative variable is a variable whose data is classified into categories or groups. The data can be of any type, so long as it represents a unique classification of that variable.

- Hypothesis test: a statistical procedure used to determine whether there is enough evidence in a sample of data to draw conclusions about a population. In a hypothesis test, two hypotheses are formulated: the null hypothesis and the alternative hypothesis.

- Null hypothesis: the null hypothesis serves as the initial assumption that any observed change or effect in the population is not statistically significant. In other words, the null hypothesis suggests that any observed differences or effects in the response of the sample are due to random variation or chance. The aim of research is to either reject or fail to reject the null hypothesis based on the sample data. Denoted H_0.

- Alternative hypothesis: the statement that contradicts the null hypothesis. It represents what the research is trying to prove or demonstrate. Acceptance of the alternative hypothesis suggests that there is a real effect, difference, or relationship. Denoted H_a.

- Significance level: the probability of rejecting the null hypothesis when it is true in a hypothesis test, commonly denoted as α.

- Expected value: represents the average value of a random variable in a probability distribution, equivalent to the mean (μ) of a random variable. Denoted as $\mathbb{E}[X]$.

- Error: the difference between the observed and the predicted values of the dependent variable generated by a regression model. Denoted as ϵ.

- Variance: measures how much each number in a dataset differs from the mean of the dataset. It quantifies the dispersion of the data points in a given dataset.

- ANOVA: acronym for analysis of variance which is a general statistical method used to analyze the differences among group means in a sample.

- Point estimate: a single value calculated from sample data that is used to estimate the corresponding population parameter. Notation is dependent on context.
- Confidence interval: a range of values derived from sample data that is likely to contain the true value of a population parameter that also provides a measure of the uncertainty associated with the point estimate known as a confidence level, which is typically 95 percent.
- P-value: a statistical measure that helps assess the evidence against a null hypothesis in a hypothesis test. It represents the probability of obtaining results as extreme as, or more extreme than, the observed results under the assumption that the null hypothesis is true. A smaller p-value suggests stronger evidence against the null hypothesis, often leading to its rejection in favor of an alternative hypothesis.
- Standard deviation: measures the average deviation of each data point from the mean in a dataset. It can be seen as a measure of the variability that exists in the data. Commonly denoted as σ.

Now that a clear understanding of necessary terminology has been established, the next portion of this section provides insight into other frequently used notations which are defined by or define some of these terms.

- β: Population parameters - representing true and unknown coefficients in a linear regression model.
  - β_0: The population parameter representing the intercept in a linear regression model. It signifies the expected value of the dependent variable when all independent variables are zero.

  - β_i; Population parameters for the slope coefficients in a linear regression model. Each β_i represents the expected change in the dependent variable for a one-unit change in the corresponding independent variable. The variables β_i are defined for all $1 \leq i \leq k$ where $i \in \mathbb{Z}$ which is the set of integers $[0, 1, 2, ..., k]$ which correspond to the set of independent variables represented in a given model.

- $\hat{\beta}$: Parameter estimates - provided by point estimates calculated from the sample data used to approximate the population parameters β.
  - $\hat{\beta}_0$: The parameter estimate for the intercept obtained from sample data. It provides an approximation of β_0 based on the observed data.
  - $\hat{\beta}_i$: Parameter estimates for the slope coefficients obtained from sample data. These estimates quantify the relationship between each independent variable and the dependent variable based on the observed sample. The variables β_i are defined for all $1 \leq i \leq k$ where $i \in \mathbb{Z}$ which is the set of integers $[0, 1, 2, ..., k]$ which correspond to the set of independent variables represented in a given model.

- Y: Response variable, representing the the true value of the dependent variable in a population.

- $\hat{Y}$: Predicted value of response variable - obtained from the linear regression model, representing the expected or estimated values of the response variable based on sample data - AKA a point estimate of the dependent variable given by a regression model.

- $\mathbb{E}[X]$: The expected value of a random variable X

- ϵ: The random error that exists between the predicted and actual values of the dependent variable. In the context of a regression model, the error

for a specific data point is known as a residual, which is given by the equation $(Y_i - \hat{Y}_i)$

- Residual Sum of Squares (RSS) - a measure of the overall goodness-of-fit of a regression model, calculated by summing the squared differences between observed and predicted values. The squaring of the error takes accounts for both positive and negative error.

- Sum of Squares Error (SSE) - the sum of squared residuals, representing the unexplained variability in the response variable not captured by the regression model. It is the same measure as RSS, however it is used to more formally denote error based on sample data.

- R^2: Coefficient of Determination (R^2) - a measure of the proportion of variability in the response variable explained by the regression model. It ranges from 0 to 1, with higher values indicating a better fitting model.

Given these frequently used terms and notations, the subsequent section of this thesis pivots towards establishing a few concrete visuals. Some of these are distribution definitions and visuals, which are meant to intuitively visualize and justify the majority of hypothesis testing in this work. The rest are visual tools which are commonly used in many statistical applications. In the context of this study, they are used to assess the validity of the model during formulation, exploration, and validation.

Relevant Visualizations and Distributions:

- Normal Distribution: the normal distribution, also known as the Gaussian distribution, is a symmetric probability distribution that is characterized by its bell-shaped curve. In a completely normal distribution, the mean, median, and mode are equal. About 68 percent of the data falls

within one standard deviation of the mean, 95 percent within two standard deviations, and 99.7 percent within three standard deviations. It is a fundamental distribution in statistics and is used as an assumption in various statistical tests.

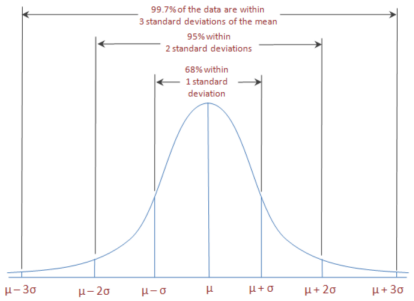

Figure 2.1: A standard normal distribution centered around the mean of the data μ with a standard deviation σ. The distribution which many assumptions that justify the construction of a regression based model are centered around.

- T-Distribution: similar to the normal distribution but has heavier tails. It is commonly used in hypothesis testing under the assumption of a normal distribution.

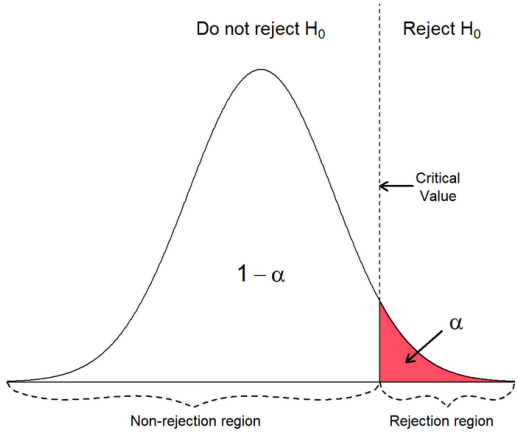

Figure 2.2: Sample T-Distribution, with a given significant level α. The area to the right and left of the T-critical values corresponds to a p-value $< \alpha$, which are the regions of significance that allows us to reject or fail to reject the null hypothesis of an F-test.

- F-Distribution: another probability distribution that arises in the context of statistical hypothesis testing for comparing variances. It is right-skewed

and its shape is determined by two sets of degrees of freedom. In the context of analysis of variance (ANOVA) and regression analysis, the F distribution is employed to assess whether the variances among group means are statistically significant.

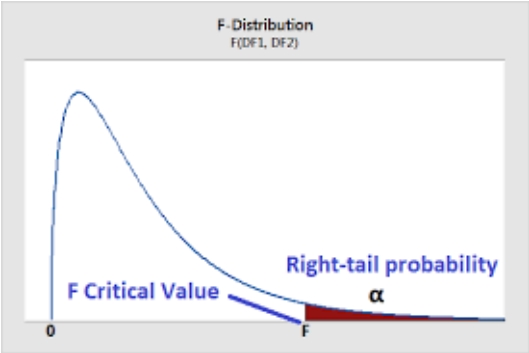

Figure 2.3: Sample F-Distribution, with a given significant level α. The area to the right of the F-critical value corresponds to a p-value $< \alpha$, which is the region of significance which allows us to reject or fail to reject the null hypothesis of an F-test.

- Histogram: a graphical representation of the distribution of a dataset. It consists of a series of bars, where each bar represents a range of values, and the height of the bar corresponds to the frequency or count of data points falling within that range. Histograms provide a visual summary of the underlying probability distribution of a continuous variable, helping to identify patterns, trends, and central tendencies in the data.

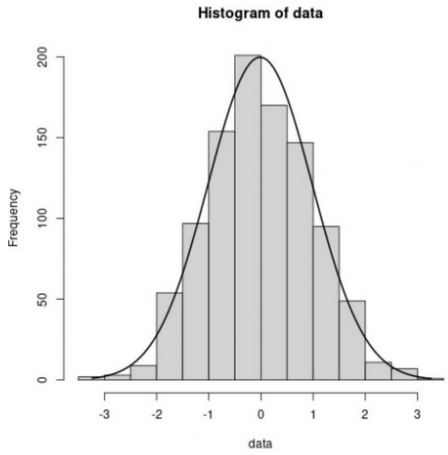

Figure 2.4: A histogram that represents a normal distribution. The histogram of a normally distributed variable should be symmetric and have an even spread centered around the mean μ.

- Q-Q plot (Quantile - Quantile plot): a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles expected from the theoretical distribution. If the points in the Q-Q plot lie approximately along a straight line, it suggests that the data is consistent with the assumed distribution. Deviations from the straight line indicate departures from the assumed distribution. Q-Q plots are valuable for checking the normality assumption and identifying potential distributional differences

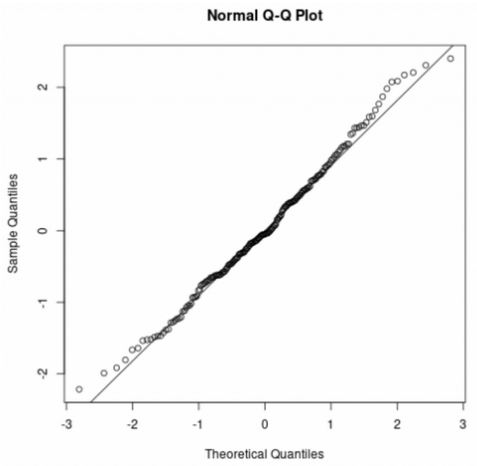

Figure 2.5: A normal Q-Q plot. A normal Q-Q plot is one that stays close to the straight line and is not skewed at the tails.

- Box plot (Box and Whisker plot): a statistical visualization that displays the distribution of a dataset. It consists of a rectangular "box" that represents the interquartile range (IQR) of the data, with a line inside the box indicating the median. Lines, or "whiskers", extend from the ends of the first and third quartiles to the minimum and maximum values within a certain range, and potential outliers may be plotted as individual points. Box plots are useful for comparing the spread and central tendency of different groups or variables as well as detecting outliers.

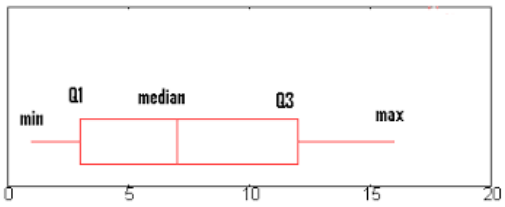

Figure 2.6: A labeled example boxplot. Any present outliers would be represented by dots extending past the ends of the plot and would not be connected to the "whiskers".

Having established a foundation in the visual representations of data distributions and tools used to assess the model during this study, it is imperative to delve into the mathematical underpinnings that form the backbone of analysis. Visualizations offer a tangible glimpse into patterns and distributions, but to

interpret and draw meaningful conclusions, a solid understanding of mathematical definitions becomes paramount. The upcoming section will define and prove key mathematical concepts that are at the core of a linear regression.

Mathematical Definitions and Equations

This section begins with an introduction to one of the most basic types of linear regression, which is a simple linear regression model, or SLR. This model can be thought of as a straight line that represents a relationship between 2 variables (hence, its name). Just like any other line, it follows the format $y = mx + b$, where an (x, y) pair denotes a single data point made up of independent x and dependent y variables. Many of the utilities of more advanced versions of this model are a result of the simple linear regression's straightforwardness. SLR can be viewed as a building block to more advanced regression techniques because it provides a fundamental understanding of how variables are related. Expansions of this model, such as multiple linear regression (the primary focus of this research) and other nonlinear regression models are better equipped for real-world scenarios.

Simple Linear regression model:

$$Y = \mathbb{E}[Y] + \epsilon \text{ , where:}$$

$$\mathbb{E}[Y] = \beta_0 + \beta_1 x$$

- β_0 : y-intercept of the regression line. The average response in the dependent variable when the independent variable has a value of 0.
- β_1 : slope of the line. Represents the amount the dependent variable increases or decreases given a 1 unit increase in the independent variable.
- x : the independent variable being used to predict the expected value of

the dependent variable.

This model serves as a basis for many principles which are absolutely crucial to understanding more advanced techniques. A couple obvious but critical observations about this model:

- By minimizing the difference between the estimation and actual value of y, our estimate of y becomes closer to the true value of y. In other words, as the random error associated with an expected value gets closer to 0, the expected value gets closer to the actual value. This improves the predictive power of the estimation.
- If we apply this logic to all points on a regression line (i.e. all of the observations in a dataset), we can assume that the resulting straight line is as close to the actual value of y as possible for all particular data points.

Thus, by minimizing the error that exists between the regression line and the actual values, we obtain the best possible estimate of our response variable, which is the goal of this type of regression. This method of minimizing errors is known as a least squares estimate and is one of the most commonly used methods of estimating the population parameters, β_i, to obtain the expected value of a given response variable (again, we do not know the true expected value of a variable because we cannot take a variable's average across every data point for an entire population, so the expected value must be estimated ($\hat{\beta}_i$)).

The following equations represent the sum of squared errors in the data (which accounts for both positive and negative error), and the least squares estimate (LSE) equation, which is used to minimize the sum of squared errors (thus, minimizing the overall error) and therefore serves as the foundational

method this research will use for predicting the expected value of the response variable:

$$\textbf{Sum of Squares (SSE):}\ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$\textbf{Least Squares Estimate (LSE):}\ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Again, hat notation is used to represent variables which are estimates. Below is a brief proof that error can be minimized via the LSE equation. It is an extension to everything discussed in the abstract thus far and is only provided as supportive knowledge.

Proof that the LSE method of minimizing error results in a regression line with optimal fit:

Consider a SLR model with one independent variable, x, and one dependent variable, y. The model can be represented as:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

Where:

- y is the observed value of the dependent variable.
- β_0 and β_1 are the parameters we want to estimate to define the best-fit line.
- x is the observed value of the independent variable.
- ε represents the error term, which captures the unexplained variability.

The goal is to find the values of β_0 and β_1 that result in the best fit to the data, i.e., the line that minimizes the error. This is defined for each data point as the difference between the observed value y and the predicted value $\hat{y}$:

$$\textbf{Error: } y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

The goal is to minimize the sum of the squared errors for ALL data points, which is known as the Residual Sum of Squares (RSS):

$$\textbf{RSS: } RSS = \sum_{i=1}^{n} (\text{Error}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that minimize the RSS are the LSE estimators. To find these estimators, we differentiate the RSS with respect to β_0 and β_1 and set the derivatives equal to zero.

$$\frac{\partial \text{RSS}}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{\partial \text{RSS}}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Solving these equations gives us the LSE estimators for β_0 and β_1, which define the best-fit line that minimizes the error.

Thus, minimizing the error via LSE results in the best fit to a regression line. A similar proof can be used to show that minimizing the error via LSE is a viable method of finding a best fit surface for a MLR model.

After proving that the LSE method of fitting a regression model is a sufficient way of obtaining a best fit, a couple important tests that will be used to repetitively improve the model can be introduced.

The first test, known as an F-test, is a type of hypothesis testing where the population variances between two groups are evaluated for significant differences. It is commonly used in the context of ANOVA to assess whether the variances of several groups are equal. In general, the F-statistic is calculated by dividing the variance between groups by the variance within groups, and the specific formula for a given F-test depends on the context of the hypothesis being tested. In general, a F-test statistic is given by the following equation. Context determines the actual values used to calculate a given F-test's statistic.

F-test statistic:

$$F_{stat} = \frac{\text{Between-group variability}}{\text{Within-group variability}}$$

In contrast, the second test, known as a t-test, is a hypothesis testing method commonly used to determine if there's a significant difference between the means of two groups. It assesses whether the sample means are signifi-

cantly different from each other, considering the variability within each group. The t-statistic is calculated by dividing the difference between group means by the standard error of this difference. The general formula for a t-test is as follows. Again, the actual values used in a given T-test's statistic is context specific, but all T-test statistics represent this idea in some manner.

T-test statistic:

$$T_{stat} = \frac{\text{Difference between group means}}{\text{Standard error of the difference}}$$

The final concepts introduced during the background of this study are the assumptions underlying any linear regression model. That is, there are a few different ideals which constitute the linearity aspect of a regression. Thus, a few assumptions must be made when using linear regression to model the relationship between a response and a predictor. These assumptions are essentially conditions that should be met before inferences can be drawn regarding the model estimates or before a model is used to make a prediction. The four primary assumptions which imply other underlying-assumptions involved with linear regression can be traced back to are the following.

1. Presence of a linear relationship: this assumption asserts that there exists a linear relationship between the independent variable and the dependent variable.

2. Independence: the assumption that the residuals are independent of one another.

3. Homoscedasticity: the assumption that the residuals have constant variance at every level of of an independent variable.

4. Normality: assumes the residuals of the model are normally distributed.

The following notation is used to denote these model assumptions, as they help articulate the underlying assumptions of the MLR model.

- **Homoscedasticity:** $\text{Var}(\varepsilon_i) = \sigma^2$
- **Expected Value:** $\mathbb{E}[\epsilon_i] = 0$
- **Independently Identically Distributed:** $\varepsilon_i \sim \text{IID}(\mu = 0, \sigma^2)$
- **Normality:** $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Moving away from the intricate mathematical formulations, definitions, and concepts introduced during this section, we delve into the practical application of these concepts in the realm of data analysis. The discussion thus far has centered on the Least Squares Estimation (LSE) method of obtaining simple linear regressions (SLR). The pivotal role of T and F tests in hypothesis testing, visual representations of relevant data distributions and tools used to examine said distributions, as well as an extensive list of key terms, notations, and mathematical concepts have been thoroughly established by this section.

As we transition into the "Data Summary" section, it is crucial to recognize that multiple linear regression (MLR) serves as a natural extension of SLR. The methods employed in our research, for constructing a comprehensive model that accommodates multiple predictors, find their roots in the LSE methodology for obtaining a line of best fit laid out in this portion of the study. Understanding the progression from SLR to MLR lays the groundwork for a nuanced exploration of the complexities inherent in analyzing the Boston housing market data set. With this in mind, it is time to briefly introduce the data set used for this work.

Chapter 3

Summary of Data

Originally published and statistically analyzed in a study titled 'Hedonic prices and the demand for clean air.' by David Harrison Jr. and Daniel L. Rubinfeld, the Boston housing market data set is represented by 14 total features across 506 observations, each of which represent a house in a unique suburb in the greater city of Boston. The data can be considered of natural origin and was collected by the U.S. census service during the 1970 census, during which the data was derived from a diverse number of sources. Additional insights into the dataset's sourcing and detailed feature descriptions can be found in the original study, which does unravel contextual nuances about the information included in the original data set. This section serves to synthesize attribute information included in that original study by Harrison and Rubinfeld; shedding light on the information represented by these variables and interpreting their potential impact on Boston housing prices in the 1970s (more information regarding feature descriptions and data sourcing can be seen in the original study[1]).

The dependent variable the regression model aims to estimate is the median value of owner-occupied homes $MEDV$ - measured in thousands of dollars. By obtaining an optimal estimation of this variable, insight can be gained on the relationships between the following 13 independent variables and the

median value of owner-occupied homes - one of the primary goals outlined in the introduction. The list of the remaining variables provided by the original dataset is comprised of the following:

- $CRIM$ - per capita crime rate by town.
- ZN - proportion of residential land zoned for lots over 25,000 square feet.
- $INDUS$ - proportion of non-retail business acres per town.
- $CHAS$ - Charles River dummy variable.
- NOX - Nitric oxides concentration (parts per 10 million).
- RM - average number of rooms per dwelling.
- AGE - proportion of owner-occupied units built prior to 1940.
- DIS - weighted distances to five Boston employment centers.
- RAD - index of accessibility to radial highways.
- TAX - full-value property-tax rate per 10,000 dollars.
- $PTRATIO$ - pupil-teacher ratio by town.
- B - the result of the equation $B = 1000 * (B_k - 0.63)^2$ where B_k is the proportion of African Americans by town.
- $LSTAT$ - lower status of the population (percentage).

As previously stated, more information regarding the origin of these features and how their values are calculated is outlined in the original study by Harrison and Rubinfeld[1]. Prior to formulating a model, some relevant information regarding these attributes and their properties needs to be explored. An example of the dataset given by the first 6 observations as well as very basic statistical summaries of the original 14 variables can be seen below.

```
> # check first observations in original data
> head(Boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio      B lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

Figure 3.1: Examples of data points in the original data set collected by Harrison and Rubinfeld

```
> # get a basic summary of initial data
> summary(Boston)
      crim                zn             indus            chas              nox               rm             age       
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      dis              rad              tax           ptratio            B              lstat            medv      
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

Figure 3.2: Basic box plot summaries of the original 14 variables

First and foremost, the summary above shows that of these 14 attributes, all but 2 are continuous numerical variables. The first of these two categorical variables ($CHAS$) can be considered a nominal or categorical variable; represented by a 1 if a given observation's tract bounds the Charles river, and a 0 otherwise. This can be seen in the summary of $CHAS$ which shows that the values for the variable are bounded by 0 and 1, and almost all observations have a value of 0 based on the variable's 3rd quartile and mean.

The second non quantitative variable (RAD) is an ordinal feature, represented by an index where a larger index denotes better accessibility to radial highways. This can similarly be confirmed by a quick glance at the variables summary, which shows that all values of RAD for a given data point are integers between 1 and 24. These qualitative features necessitate one particularly important preprocessing step that needed to be performed before generating an initial model - which was to encode them as factors.

One important concept that needs to be introduced is the interpretation of the parameter estimates which represent our variables. In general for numeric variables, the parameter estimates represent the average resulting change in the response variable given a 1 unit increase in that variable while holding all others constant. For instance, the parameter estimate for the variable RM (the average number of rooms per dwelling) represents the average change in the response variable $MEDV$ (the median value of owner-occupied homes, measured in thousands of dollars) given a 1 unit increase in RM while holding all other variables constant. For example: if the parameter estimate, $\hat{\beta}$, for the variable RM is 5; that would imply that for every additional room in a dwelling, the median value of an owner-occupied home increases by 5000 dollars, holding all other variables constant. This interpretation is applicable to any quantitative variable, however it does not apply to qualitative variables.

The parameter estimate for a qualitative variable can be represented as estimates of dummy variables that correspond to the different categories of that qualitative variable. They denote the difference in average response between one level of the variable and the reference category. For example, in the context of the Boston housing market data, a parameter estimate for the RAD variable where $RAD = 2$ represents the difference in the average response in the dependent variable between the values of $RAD = 1$ (the reference category) and $RAD = 2$ while holding the other categorical variable $CHAS$ at a consistent level and adjusting for the effect of the other variables present in the model. This interpretation can be employed on all levels of both qualitative features of the data set ($CHAS$ and RAD).

The concept of interpreting parameter estimates will be discussed in fur-

ther detail after model validation when drawing conclusions about the optimized model. With a comprehensive overview of the Boston housing market dataset and some generalized methods of generating interpretations for parameter estimates within the context of this data, the research now switches gears towards the crucial phase of data preprocessing. The data summary has laid the foundation by presenting key features and their potential impact on housing prices. As preprocessing begins, the objective is to refine and prepare the dataset for the application of MLR methods. This involves addressing missing values, exploring a general first order linear model, ensuring the data adheres to the assumptions necessary for regression analysis, investigating outliers, and splitting the data into training and testing sets for model validation. The insights gathered from the data summary will guide the preprocessing strategies used in this work, allowing for a more nuanced understanding of the intricate relationships between variables and facilitating a more accurate modeling procedure.

Chapter 4

Data Exploration and Preprocessing

Exploration of Data and a General First Order Multi Linear Model

A general first order multiple linear model based on the LSE of $\hat{y}$, where $y = MEDV$ is given by the following equation:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{CRIM} + \beta_2 \cdot \text{ZN} + \beta_3 \cdot \text{INDUS} + \beta_4 \cdot \text{NOX} + \beta_5 \cdot \text{RM} + \beta_6 \cdot \text{AGE} + \beta_7 \cdot \text{DIS} + \beta_8 \cdot \text{TAX} + \beta_9 \cdot \text{PTRATIO} + \beta_{10} \cdot \text{B} + \beta_{11} \cdot \text{LSTAT} + \beta_{12} \cdot \text{RAD} + \beta_{13} \cdot \text{CHAS}$$

This model fully encompasses the information provided by the features of the original Boston housing market data set. The respective variable names for all β parameter estimates can be seen in the R output below, which were calculated using all observations of the data set. Using the lm() function from the base R package to construct the model above, a basic summary of the parameter estimates and an ANOVA table were generated. This initial MLR model was used as a reference point during research, and served as a baseline model which could be improved upon. It also provided some important basic statistics that needed to be analyzed before further research could be conducted. A summary of this model on the original data set that was generated

using R is provided below.

```
Call:
lm(formula = medv ~ crim + zn + indus + nox + rm + age + dis +
    tax + ptratio + B + lstat + rad + chas, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
chas         2.687e+00  8.616e-01   3.118 0.001925 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Figure 4.1: Summary of initial general multiple linear model

One notable statistic whose interpretation takes precedence over the rest of the analysis of this model is the global F-test statistic which can be seen in the summary output. As seen above, the F-test stat is 108.1. The F-test represented by this statistic, known as a global F-test, is given by the equation below and provides the framework to develop the following null and alternative hypothesis:

$$F_{\textbf{stat}} = \frac{\textbf{MSM}}{\textbf{SSE/dF Error}}$$

$$MSM \text{ (Mean Square Model)} = \frac{\text{SSM}}{\text{df Model}}$$

$$SSE \text{ (Sum of Squares Error)}$$

Null and alternative hypothesis for global F-test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_a : \text{at least one of } \beta_1 ... \beta_k \neq 0$$

The null hypothesis for the global F test states that all of the parameter coefficient estimates in the model equal 0, meaning that the proven value of the dependent variable is a function of only the intercept of the fitted regression surface and the random error involved in the model. The alternative implies that at least one of the parameter coefficients is not 0. Therefore, failure to reject the null hypothesis indicates that the model is not useful at all, which is why exploring this summary statistic is so relevant at the beginning of research. The F-test statistic corresponds to a p-value $< 2.2 * 10^{-16}$, meaning we fail to reject the alternative hypothesis with a probability of almost 100 percent and can therefore conclude that at least one of the parameter estimates of this first order linear model is statistically significant enough to contribute to the expected value of the dependent variable $MEDV$.

Next, we turn our attention to the multiple coefficient of determination and the adjusted multiple coefficient of determination, R^2. The output summary provides us with a R^2 value of 0.7406 and a R_a^2 value of 0.7338. These statistics both represent the proportion of variability in the response variable explained by the model (where $R^2 = 1$ represents a perfectly fit model and $R^2 = 0$ represents a complete lack of fit in the model). The R_a^2 stat is always less than or equal to R^2, as it takes into account the sample size n and the number of parameters in the model k. Thus, the R_a^2 statistic informs us that only about 73 percent of the variance of the response variable is explained by this model, meaning that there is certainly room for improvement in terms of the performance of this general first order model. A closer examination of the individual variables in the model will be conducted during model formulation

and at the end of research. For now, some initial diagnostics to properly preprocess the data need to be considered.

First off, the normality assumption ($\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$) applies to the response variable. To conduct a brief analysis of these assumptions during this portion of the research, multiple graphs to visualize the data were used to come to a conclusive decision regarding the normality of the data.

To check the validity of these assumptions on the general first order linear model, a Q-Q plot and a histogram were created and analyzed for each of the numeric continuous variables in the dataset. The figures below showcase these graphs, which prove that our assumptions are not fully satisfied. Ideally, the data points would be close to the straight line on the Q-Q plots and not 'tail' out at the ends. Data points that are far from the line indicate potential outliers which will be further analyzed. Similarly, the histogram's distributions should be centered around the variables' respective averages and be symmetric, with no gaps, and an equal spread on both sides. A quick glance at the R output below proves that the data is not normally distributed, as none of the Q-Q plots or histograms follow these guidelines which define normally distributed variables. One thing to note is that the Q-Q plot and histogram of the response variable $MEDV$ do appear to show a normally distributed variable. This is great because the model assumptions regarding the residuals otherwise constitute a normalization process based on the Q-Q plots and histograms. However, since the response variable is already somewhat normalized, we can conclude that a transformation to the response will most likely normalize the residuals.

The graphs indicate that either a normalization process needs to be applied

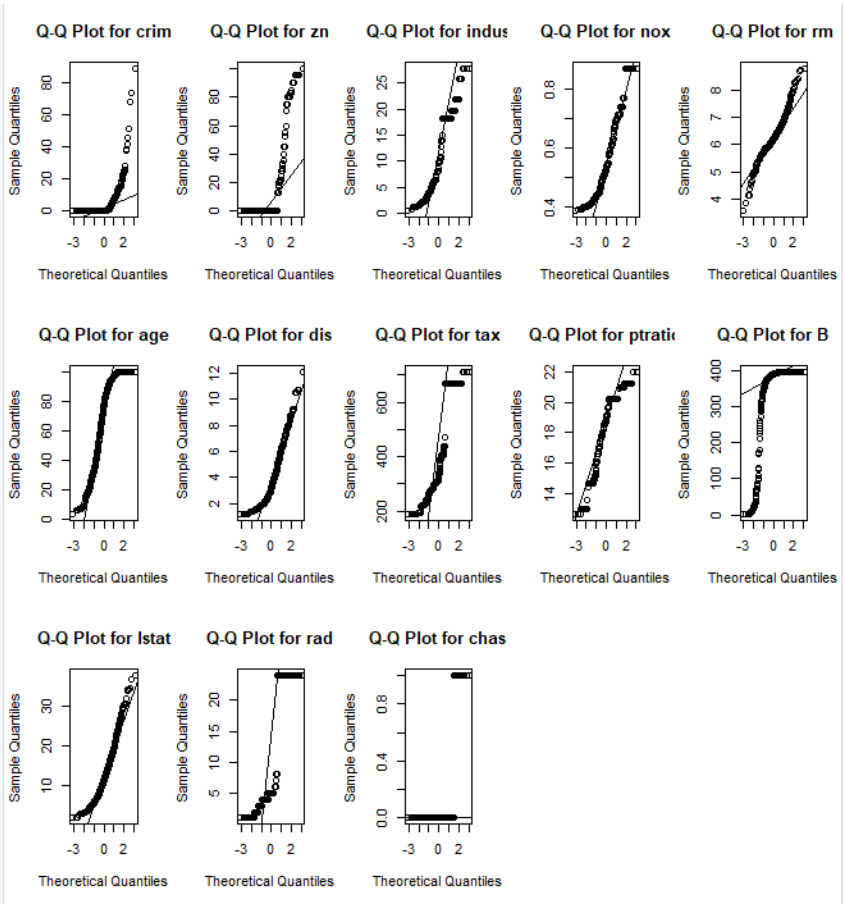

Figure 4.2: Q-Q plots of each variable in original data set

to the independent variables or a transformation of the response variable needs to occur. For the purposes of this research, these methods of addressing the assumption violation will be performed in the model validation portion of this thesis. This is a deliberate decision because the normalization of these variables would complicate their interpretations on the response variables. There are also constraints on the data which indicate that normalization would not be generally useful for some variables, and normalizing a subset of the independent variables typically introduces bias into the model. For instance, the Q-Q plot and histogram for the variables CRIM and ZN indicate that a significant number of data points have a value of 0 for these variables, meaning if we normalize the data it will affect the interpretation of these 0 values. Furthermore, many of the variables such as B, PTRATIO, AGE, and INDUS

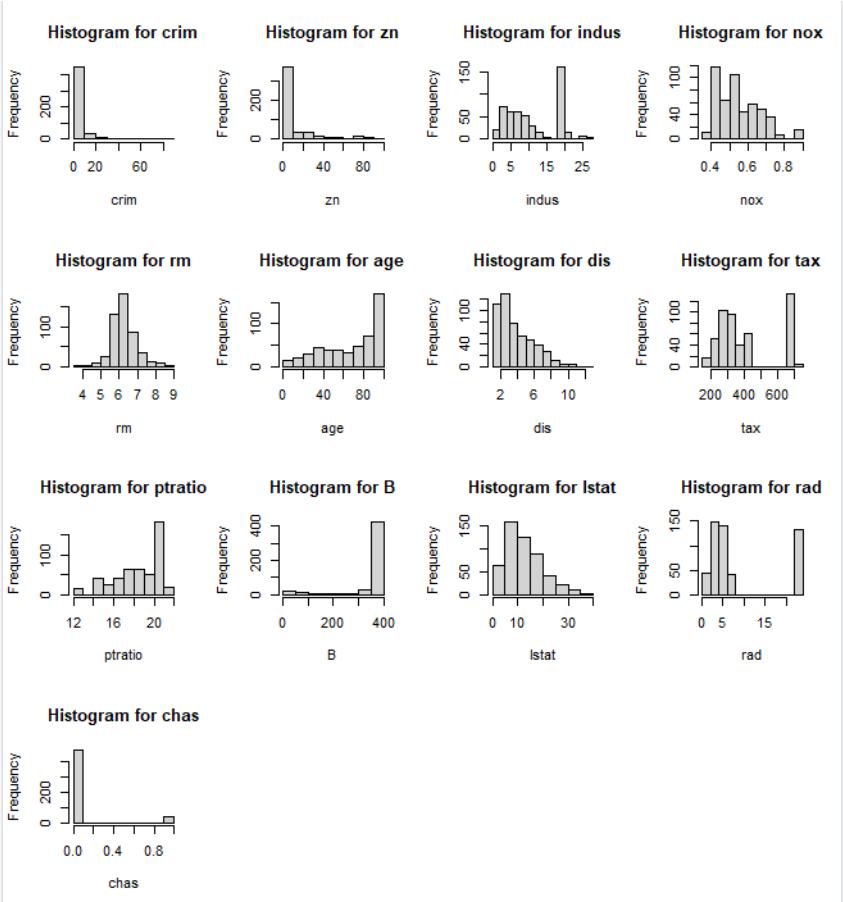

Figure 4.3: Histograms of each variable in original data set

represent proportions, which are already on a scale and if normalized will take away from their context in our research, as they are some of the variables whose relationships are intended to be explored in the most depth. The alternative method of transforming the response variable will be used to try and satisfy the normality assumptions during model validation. For now, we continue to preprocess the data in hopes that handling outliers, introducing categorical data, and taking care of missing values will improve our main effects model and provide some improvement regarding the normality of the data.

Preprocessing

As previously stated, this data set is extraordinarily 'clean' relative to most others. Thus, a lot of the data preprocessing that would typically be involved in creating a regression model was not performed during this research. However, it is still necessary to perform some initial diagnostics before attempting to improve the general first order model.

One thing to be noted which is evidenced by the following R output, the data set did not include any missing values for any of the 14 features, meaning that none of the data in the housing market data set was removed or imputed prior to modeling. This is a positive attribute about the Boston housing market data set, because usually missing values need to be accounted for and estimated, which can add discrepancies and ambiguity to the interpreted results and predictive capabilities of a model. The following check in R was used to verify this.

```
> #check for missing values
> any(is.na(Boston))
[1] FALSE
```

Figure 4.4: Proof that there are no missing values present in the data

Due to the absence of NA values, it was possible to directly split the original Boston housing market data set into training and test sets. There are many possible ways to optimize the splitting of data into these subsets which are beyond the scope of this research. For the purpose of training this model, a simple 80-percent/20-percent split of the data into training/testing sets was performed. This split ratio is a somewhat standardized practice for training models that do no have specific constraints which constitute more specialized splitting ratios. The following R output evidences this split of the data set. As seen below, the resulting training/testing data sets contain 407 and 99

observations respectively, which roughly represents a 80/20 split of the original 506 observations.

```
> training_data <- Boston[index, ]
> testing_data <- Boston[-index, ]
> str(training_data)
'data.frame':   407 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.03237 0.06905 0.08829 ...
 $ zn     : num  18 0 0 0 12.5 12.5 12.5 12.5 12.5 0 ...
 $ indus  : num  2.31 7.07 2.18 2.18 7.87 7.87 7.87 7.87 7.87 8.14 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.458 0.458 0.524 0.524 0.524 0.524 0.524 0.538 ...
 $ rm     : num  6.58 6.42 7 7.15 6.01 ...
 $ age    : num  65.2 78.9 45.8 54.2 66.6 96.1 85.9 82.9 39 56.5 ...
 $ dis    : num  4.09 4.97 6.06 6.06 5.56 ...
 $ rad    : int  1 2 3 3 5 5 5 5 5 4 ...
 $ tax    : num  296 242 222 222 311 311 311 311 311 307 ...
 $ ptratio: num  15.3 17.8 18.7 18.7 15.2 15.2 15.2 15.2 15.2 21 ...
 $ B      : num  397 397 395 397 396 ...
 $ lstat  : num  4.98 9.14 2.94 5.33 12.43 ...
 $ medv   : num  24 21.6 33.4 36.2 22.9 27.1 18.9 18.9 21.7 19.9 ...
> str(testing_data)
'data.frame':   99 obs. of  14 variables:
 $ crim   : num  0.0273 0.0299 0.2112 0.2249 0.6298 ...
 $ zn     : num  0 0 12.5 12.5 0 0 0 0 0 75 ...
 $ indus  : num  7.07 2.18 7.87 7.87 8.14 8.14 8.14 8.14 5.96 2.95 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.469 0.458 0.524 0.524 0.538 0.538 0.538 0.538 0.499 0.428 ...
 $ rm     : num  7.18 6.43 5.63 6.38 5.95 ...
 $ age    : num  61.1 58.7 100 94.3 61.8 84.5 94.1 100 68.2 15.8 ...
 $ dis    : num  4.97 6.06 6.08 6.35 4.71 ...
 $ rad    : int  2 3 5 5 4 4 4 4 5 3 ...
 $ tax    : num  242 222 311 311 307 307 307 307 279 252 ...
 $ ptratio: num  17.8 18.7 15.2 15.2 21 21 21 21 19.2 18.3 ...
 $ B      : num  393 394 387 393 397 ...
 $ lstat  : num  4.03 5.21 29.93 20.45 8.26 ...
 $ medv   : num  34.7 28.7 16.5 15 20.4 18.2 12.7 14.5 18.9 34.9 ...
>
> # Check the dimensions of the training and testing sets
> cat("Training Set Dimensions:", dim(training_data), "\n")
Training Set Dimensions: 407 14
> cat("Testing Set Dimensions:", dim(testing_data), "\n")
Testing Set Dimensions: 99 14
```

Figure 4.5: Data type summaries and dimensions of training/testing sets after 80/20 split of original data.

Note that the model being explored during this portion of research still misrepresents the data types for both qualitative variables, which can be seen in figure 4 (showing that the data types of RAD and $CHAS$ are being interpreted as numeric variables represented by integers). This begins the final portion of data preprocessing - encoding qualitative variables as integers. The methodology to encode RAD, known as label encoding, is a technique commonly employed in machine learning when dealing with categorical variables. In the context of the Boston housing dataset, the variables RAD (index of accessibility to radial highways) and CHAS (Charles River dummy variable) are categorical in nature, representing discrete categories or levels. To facilitate the incorporation of these variables into regression models, they are often

encoded as factors in R.

Label encoding for the RAD variable is particularly beneficial due to the ordered nature of its values. RAD represents different levels of accessibility to radial highways, ranging from 1 to 24. Assigning numerical labels to these levels reflects the inherent order and increasing magnitude of accessibility. This encoding captures the ordinal relationship between the categories, allowing the model to interpret the varying degrees of radial highway accessibility. Consequently, the multilinear regression model can leverage this ordered encoding to discern the impact of changing RAD levels on the predicted outcome, providing a more nuanced understanding of the relationship between accessibility and housing prices in the Boston dataset. By incorporating label encoding, the model can navigate the categorical nature of RAD while preserving the meaningful information embedded in the ordered levels of accessibility.

```
> str(training1)
'data.frame':   407 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.03237 0.06905 0.08829 ...
 $ zn     : num  18 0 0 0 12.5 12.5 12.5 12.5 12.5 0 ...
 $ indus  : num  2.31 7.07 2.18 2.18 7.87 7.87 7.87 7.87 7.87 8.14 ...
 $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ nox    : num  0.538 0.469 0.458 0.458 0.524 0.524 0.524 0.524 0.524 0.538 ...
 $ rm     : num  6.58 6.42 7 7.15 6.01 ...
 $ age    : num  65.2 78.9 45.8 54.2 66.6 96.1 85.9 82.9 39 56.5 ...
 $ dis    : num  4.09 4.97 6.06 6.06 5.56 ...
 $ rad    : Factor w/ 9 levels "1","2","3","4",..: 1 2 3 3 5 5 5 5 5 4 ...
 $ tax    : num  296 242 222 222 311 311 311 311 311 307 ...
 $ ptratio: num  15.3 17.8 18.7 18.7 15.2 15.2 15.2 15.2 15.2 21 ...
 $ B      : num  397 397 395 397 396 ...
 $ lstat  : num  4.98 9.14 2.94 5.33 12.43 ...
 $ medv   : num  24 21.6 33.4 36.2 22.9 27.1 18.9 18.9 21.7 19.9 ...
```

Figure 4.6: Proof of encoding the qualitative variables as factors in the split data sets. The training1 datatset is the dataset that is used during model formulation and validation.

From this point forward, any statistical analysis of the model is being performed on the training dataset seen above. After the MLR model is optimized, the testing data will be used to predict the expected value of $MEDV$ which will then be compared to the actual values of $MEDV$ found in the test data. Having successfully preprocessed the data, including label encoding for categorical variables, we now turn our attention to the formulation of the MLR

model.

Chapter 5

Model Formulation

The beginning of model formulation necessitates the introduction of the concept of overfitting a model. Model overfitting is a common downfall in the realm of statistical modeling. It occurs when a model becomes excessively complex and starts fitting not just the underlying patterns in the data but also the noise. In the context of this research, the pursuit of capturing every nuance within the training dataset can cause an overfit model, which tends to lose its generalization capability to new, unseen data (which will be provided to the model during validation.

A crucial principle in model formulation is the delicate trade-off between model complexity and generalization. While a complex model may achieve remarkable accuracy on the training data, it also is likely to struggle when exposed to new observations. Simpler models, on the other hand, often generalize better to unseen data by capturing the underlying trends without being overly influenced by noise. Striking the right balance is essential.

The statistical philosophy of simplicity being sufficient (or even preferential to a more complicated model) to explain the observed data advocates for choosing the simplest model that adequately represents the underlying patterns in the data.This simplicity not only enhances model interpretability but

also prevents overfitting.

In summary, an overfit model may lead to misleading conclusions, especially in scenarios where the model needs to make predictions on new, unseen data. During the model formulation process, careful consideration must be given to factors such as feature selection, regularization techniques, and cross-validation to prevent overfitting. By prioritizing simplicity and generalization, a well-crafted model not only avoids the pitfalls of overfitting but also becomes a reliable tool for making accurate predictions in diverse test settings. These concepts will be explained more thoroughly after the first order numeric main effects portion of the model is finalized.

Simplification of Numeric Portion of Main Effects Model

The beginning of model formulation introduces the process of model selection. It will be used multiple times throughout formulation to determine the most significant variables in the model. In this initial step of formulation we apply hypothesis testing to the parameter estimates and apply model selection strategies to the main effects model, given by the following equation:

Full initial main effects model:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{CRIM} + \beta_2 \cdot \text{ZN} + \beta_3 \cdot \text{INDUS} + \beta_4 \cdot \text{NOX} + \beta_5 \cdot \text{RM} + \beta_6 \cdot \text{AGE} + \beta_7 \cdot \text{DIS} + \beta_8 \cdot \text{TAX} + \beta_9 \cdot \text{PTRATIO} + \beta_{10} \cdot \text{B} + \beta_{11} \cdot \text{LSTAT} + \beta_{12} \cdot \text{RAD} + \beta_{13} \cdot \text{CHAS}$$

Obviously, this main effects model takes into account everything provided by the dataset, which is why the number of variables included in the model is so high. Some of these variables provide redundant information or are not contributing significantly to the model. In addition, since the variables are of differing data types, their influence on the response variable needs to be analyzed separately and with different methodologies. First, the numeric

portions of the The goal of this portion of model formulation is to reduce the number of variables in the model which will help avoid overfitting the model and simplify the analysis performed in further steps. First, we evaluate the validity of a null and alternative hypothesis for each individual quantitative parameter estimate, with an individual t-test. The hypothesis statements below reflect the implications of the T-test on each variable of the numeric portion of the full initial main effects model which follows as:

Numeric portion of full initial main effects model:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{CRIM} + \beta_2 \cdot \text{ZN} + \beta_3 \cdot \text{INDUS} + \beta_4 \cdot \text{NOX} + \beta_5 \cdot \text{RM} + \beta_6 \cdot \text{AGE} + \beta_7 \cdot \text{DIS} + \beta_8 \cdot \text{TAX} + \beta_9 \cdot \text{PTRATIO} + \beta_{10} \cdot \text{B} + \beta_{11} \cdot \text{LSTAT}$$

T-test null and alternative hypothesis for each individual parameter estimate:

$$H_0 : \beta_0 = 0, \beta_1 = 0, ... = \beta_{11} = 0$$
$$H_a : \beta_0 \neq 0, \beta_1 \neq 0, ... = \beta_{11} \neq 0$$

This hypothesis test was conducted on all of the numeric parameter estimates using the output of the summary() function from the base package of R. The output below provides the T-test statistics for all these estimates as well as their respective p-values:

This output of parameter estimates provides a basic foundation of knowledge which will be used to validate or invalidate the model selection processes used proceeding this step. The R output above describes the effects of the parameter estimates on the response variable. Any estimate that has does not have a dot, or one or more stars, to the right of its respective p-value represents a parameter that fails to reject the null hypothesis, indicating that the variable does not contribute significantly to predicting the response, and should therefore be removed from the model. In summary, the individual t-

```
Residuals:
    Min      1Q  Median      3Q     Max
-12.943  -2.799  -0.569   1.675  26.808

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.200593   5.558592   5.613 3.75e-08 ***
crim         -0.053601   0.039005  -1.374  0.17016
zn            0.033247   0.016403   2.027  0.04335 *
indus        -0.065332   0.069480  -0.940  0.34764
nox         -14.382263   4.438061  -3.241  0.00129 **
rm            3.848592   0.459839   8.369 1.01e-15 ***
age           0.005968   0.015946   0.374  0.70839
dis          -1.426272   0.237659  -6.001 4.43e-09 ***
tax           0.001870   0.002694   0.694  0.48792
ptratio      -0.881040   0.152060  -5.794 1.41e-08 ***
B             0.009282   0.003017   3.076  0.00224 **
lstat        -0.561580   0.061000  -9.206  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.95 on 395 degrees of freedom
Multiple R-squared:  0.7164,    Adjusted R-squared:  0.7085
F-statistic: 90.69 on 11 and 395 DF,  p-value: < 2.2e-16
```

Figure 5.1: Output of summary() function for numeric portion of full main effects model

test statistics of these parameters do not have probabilities smaller than a specified significance level (in this case and most cases, a significance level of 0.05 is used) indicating the probability that the t-test statistic is greater than a t-critical value (which is based on the significance level as well as the degrees of freedom in the model) is less than 95 percent. This implies the t-test does not provide sufficient evidence to reject the null hypothesis. The parameters whose estimates should not be included in the main effects model based on their individual t-tests are CRIM, INDUS, AGE, and TAX. Referring back to the interpretation of the parameter estimate coefficients and the results of their individual t-tests allows us to draw the following conclusions regarding these four variables:

- A one unit increase in per capita crime rate by town does not provide any significant contribution to the estimation of the median value of owner occupied homes in Boston.
- A one unit increase in the proportion of non-retail business acres per town does not provide any significant contribution to the estimation of

the median value of owner occupied homes in Boston.

- A one unit increase in the proportion of owner-occupied units built prior to 1940 does not provide any significant contribution to the estimation of the median value of owner occupied homes in Boston.
- A one unit increase in the full-value property-tax rate per 10,000 dollars does not provide any significant contribution to the estimation of the median value of owner occupied homes in Boston.

To verify the results of the individual t-tests, 2 model selection processes are applied to the numeric variables in the main effects model. These selection processes are essentially automated search methods, and are known as backwards and stepwise selections. They are not guaranteed to give us the best model and are also not guaranteed to generate the same model. Thus, we use them in conjunction with the aforementioned individual t-tests and each other to verify that the variables that will be removed from the quantitative portion of the main effects model are indeed the best features of the dataset to exclude from the model. Before analyzing the results of these processes and comparing them to those generated from the individual t-tests, a quick summary of the model selection processes is outlined to improve the interpretability of the results provided from the search methods.

Each process can be described by 4 steps. To perform backwards selection or 'elimination', the model with all predictor variables of interest is taken into consideration. For each term in the model, a p-value is generated and the term with the largest is removed from the model. The model is then refitted without the eliminated predictor, and the process is repeated until all of the remaining predictors are statistically significant (their p-values are less than the chosen significance level). In contrast, forward selection begins with an

empty model, and predictors are evaluated against the response variable one at a time. The variable with the smallest p-value is inserted into the model, which is then refitted against the response variable. This process repeats for each remaining predictor one at a time until none of the remaining predictors that have not been inserted into the model are significant. Similarly, stepwise selection starts with an empty model. Forward selection is applied to the model at each step and a re-evaluation of the previously inserted predictors is performed. If a previously added predictor has become insignificant, remove it and add it back to the list of potential variables. These model selection processes will be performed many times over the formulation and fitting of the final model, so a general understanding of the methodology they invoke is crucial to interpreting the results of the processes which will continuously help improve the model. With this in mind, observe the following R output; which represents the resulting lists of predictor variables generated from using these automated search methods on the numeric portion of the main effects model.

```
Step:  AIC=1308.55
medv ~ zn + nox + rm + dis + ptratio + B + lstat

          Df Sum of Sq     RSS    AIC
<none>                  9745.9 1308.5
- zn       1    103.45  9849.4 1310.8
- B        1    285.91 10031.8 1318.3
- nox      1    393.37 10139.3 1322.7
- dis      1    984.41 10730.3 1345.7
- ptratio  1   1126.71 10872.6 1351.1
- rm       1   1898.79 11644.7 1379.0
- lstat    1   2659.93 12405.8 1404.8
```

Figure 5.2: Results of backwards elimination on numeric portion of initial full main effects

As seen in the above R output, of the three selection methods, backwards and stepwise resulted in a unanimous set of predictor variables, which further validates the exclusion of the predictor variables provided by the 4 parameter estimates who all failed their individual t-tests. This solidifies the quantitative portion of the initial main effects model, which is the following:

```
Step:  AIC=1308.55
medv ~ zn + nox + rm + dis + ptratio + B + lstat

          Df Sum of Sq     RSS    AIC
<none>                  9745.9 1308.5
+ crim     1     37.91  9708.0 1309.0
+ indus    1     14.29  9731.6 1310.0
+ age      1      4.46  9741.5 1310.4
+ tax      1      0.01  9745.9 1310.5
- zn       1    103.45  9849.4 1310.8
- B        1    285.91 10031.8 1318.3
- nox      1    393.37 10139.3 1322.7
- dis      1    984.41 10730.3 1345.7
- ptratio  1   1126.71 10872.6 1351.1
- rm       1   1898.79 11644.7 1379.0
- lstat    1   2659.93 12405.8 1404.8
```

Figure 5.3: Results of stepwise elimination on numeric portion of initial full main effects

Numeric portion of main effects model:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT}$$

Simplification of Qualitative Portion of Main Effects Model

This reduction of quantitative numeric variables has some negligible effects on the predictive capabilities of the model. As previously mentioned, the main benefit derived from variable reduction is that it improves interpretability, simplicity, and avoids overfitting the model. However, this reduction has only been applied to the quantitative variables of the model. A reintroduction of the categorical variables back into the model results in another substantial layer of complexity. Therefore, it is necessary to decide if this added complexity that will be present in the resulting model is statistically significant enough to justify the predictive advantage the model would gain from the additional variable. Because the complexity introduced to the model by a qualitative feature is M-1 variables for an M level categorical variable, a test must be used that combines information across these M-1 levels that pertain to a sin-

gle categorical variable. This is a task that is most optimally handled by a partial F-test. In this case, the reduced model is the quantitative portion of the main effects model with the 7 independent variables seen above. The complete model is one where the categorical variables are added to the reduced model. Because there are two categorical variables, this results in 4 separate models (the reduced model, the reduced model with one categorical variable, the reduced model with the other categorical variable, and the reduced model with both categorical variables - the full model). These models are given by the following equations:

Reduced model:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT}$$

Reduced model with RAD:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT} + \beta_8 \cdot \text{RAD}$$

Reduced model with CHAS:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT} + \beta_8 \cdot \text{CHAS}$$

Full model:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT} + \beta_8 \cdot \text{RAD} + \beta_9 \cdot \text{CHAS}$$

The partial F-test for these different nested models were compared in R. The following ANOVA results depict the partial F-test statistics and p-value. Evidenced by these outputs, both the reduced model with CHAS and the reduced model with RAD had a partial F-test statistic with a P-value $< \alpha$, meaning

we reject the null hypothesis and conclude that when independently removed from the full numeric main effects portion of the model, the coefficient of RAD AND CHAS are both non-zero. Thus, we conclude that the removal of both coefficients from the model has a statistically significant impact on the response variable, and they should therefore be included in the model.

```
> anova(onlyNumsF, modelWrad)
Analysis of Variance Table

Model 1: medv ~ zn + nox + rm + dis + ptratio + B + lstat
Model 2: medv ~ zn + nox + rm + dis + ptratio + B + lstat + rad
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    399 9745.9
2    391 9051.5  8    694.44 3.7497 0.0003018 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(onlyNumsF, modelWchas)
Analysis of Variance Table

Model 1: medv ~ zn + nox + rm + dis + ptratio + B + lstat
Model 2: medv ~ zn + nox + rm + dis + ptratio + B + lstat + chas
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    399 9745.9
2    398 9554.1  1    191.84 7.9916 0.004937 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(onlyNumsF, fullModel)
Analysis of Variance Table

Model 1: medv ~ zn + nox + rm + dis + ptratio + B + lstat
Model 2: medv ~ zn + nox + rm + dis + ptratio + B + lstat + rad + chas
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    399 9745.9
2    390 8907.7  9    838.26 4.0779 5.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.4: Partial F-tests of reduced, partial, and full models

Evidenced by the above ANOVA outputs from R, both the reduced model with CHAS and the reduced model with RAD had a partial F-test statistic with a P-value $< \alpha$, meaning we reject the null hypothesis and conclude that when independently removed from the full numeric main effects portion of the model, the coefficient of RAD AND CHAS are both non-zero. Thus, we conclude that the removal of both coefficients from the model has a statistically significant impact on the response variable, and they should therefore be included in the model.

This leaves us with the following finalized main effects model:

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT} + \beta_8 \cdot \text{RAD} + \beta_9 \cdot \text{CHAS}$$

Checking for Interactions and Potential Higher Order Models:

The solidified main effects portion of the model represents what can be considered a first order model of the Boston housing data. That is, the model does not include terms of a higher degree than 1 (variables which are raised to a power higher than 1). Models of this type are simplistic in nature and easier to interpret than higher order models; which introduce complexities that can obscure the fundamental insights sought in regression analysis. For instance, introducing quadratic or interaction terms (the two most commonly explored higher order terms in a MLR) may lead to multicollinearity - where predictor variables become highly correlated with one another. This challenges the model's stability and invalidates its sufficiency. Autocorrelation issues may arise as higher-order terms complicate the relationships between variables, potentially distorting the model's predictive accuracy. Further, the risk of overfitting increases, as the model may adapt too closely to the training data, compromising its ability to generalize to new, unseen data. These issues and their related concepts will be further discussed during the proceeding finalization of the above model and during model validation, but for now the main point of focus is avoiding them by keeping the model as simplistic as possible. Including interactions and higher order terms in the model is a very open ended point of research that requires an extreme amount of model refinement, which is beyond the scope of this thesis. However, exploring the possibilities of a higher order model on a surface level should be considered for potential

future research pursuits.

As far as potential interaction terms are concerned, determining whether to include an interaction term in a model is a critical decision, and it requires careful consideration of several factors. Interaction terms are introduced to account for the joint effect of two or more variables, allowing for a more nuanced understanding of their combined impact on the response variable. However, including interaction terms comes with challenges and considerations.

First and foremost, the inclusion of an interaction term should align with an underlying theory or subject-matter being investigated. It is necessary to have a plausible rationale or hypothesis supporting the belief that the relationship between certain variables is not additive but rather influenced by their joint effect. Without a theoretical basis, adding interaction terms does little more than lead to overfitting a model that lacks interpret-ability. Since the scope of this research is simply to optimize a MLR model and use definitive results from that model to interpret the existing relationships to gain a baseline understanding of their effects on the response variable, no interaction terms will be included in the model which gets validated with the test data. In other words, this thesis would serve as a baseline for further research to test potential interactions of interest - which could hypothetically be found based on the interpretations of the relationships discovered from the main effects model formulated in this research.

Secondly, multicollinearity is a concern when introducing interaction terms. High correlation between the interacting variables can destabilize the model and make it challenging to isolate the unique contribution of each variable. With a model as large as the one deemed significant up to this point in re-

search (a model with a total of 9 independent variables), it does not make sense to go through each potential interaction (since we have no hypothesis generated to single out specific interactions of interest) and test the variance inflation introduced by the addition of the interaction in the model, as following through with this process would introduce a seemingly endless amount of model refinement. This is the aformentioned surface level idea which was briefly explored during model formulation in pursuit of uncovering motivation for further research using R. Unfortunately, the effort was in vein, as there were too many significant interactions to investigate without having proper supporting knowledge to suspect an interaction exists.

The established main effects component of the model embodies a first-order approach to the Boston housing data, abstaining from the inclusion of terms exceeding the first degree. This design champions simplicity, offering enhanced interpretability compared to more complex higher-order models. With respect to the other types of higher order terms that would be appropriate to research at this point in the model building process - i.e. testing polynomial terms - the most sensible first step is the introduction of quadratic terms, representing variables squared. These terms are likely to introduce potential pitfalls that can compromise the integrity of the regression analysis. Similar to interaction terms, quadratic terms may lead to multicollinearity and overfitting the model, which challenges the model's stability and renders it vulnerable to instabilities which are imminent when the model is fed new data. Another point of avoiding quadratic terms is the even more complex interpretation of parameter estimates. Understanding the impact of quadratic terms on the response variable becomes challenging, as the relationships are no longer linear and straightforward at all. Unlike interaction terms they do not offer valuable insight about the joint effect of two variables on the response. Thus the chal-

lenges invoked by the introduction of such quadratics are even less justified in terms of research efforts than those of interaction terms. In the context of this work, the decision to avoid quadratic terms is deliberate for similar reasons as interactions and all higher order terms. While they may capture certain nonlinear relationships, introducing higher order terms without a compelling theoretical basis can lead to a loss of interpretability and the risk of fitting noise rather than genuine patterns in the data. Given the study's focus of optimizing a multiple linear regression (MLR) model for practical insights, the simplicity of a first-order model is favored. The avoidance of quadratic terms aligns with the overarching goal of creating a model that not only performs well in terms of predictive accuracy but also remains interpretable and resistant to the common challenges associated with higher-order complexities.

In conclusion, the formulation of this model has been a meticulous process, guided by variable selection techniques such as t-tests, automated selection methods, and partial F-tests. These approaches allowed the construction of a robust main effects model, emphasizing simplicity and interpretability. The decision to exclude higher-order terms was deliberate and based on the recognition that the introduction of such terms could potentially obscure the clarity of the initial model. The pursuit of simplicity serves as a crucial principle in the model formulation process this work emphasizes. Given the intricate challenges that can arise from the inclusion of higher-order terms, especially without a compelling thesis to justify their incorporation, we have prioritized a parsimonious model that strikes a balance between accuracy and interpretability. This foundational model not only enhances our understanding of the underlying relationships but also sets the stage for subsequent analyses. The conclusions that will be drawn from the first-order model during validation will provide a valuable baseline for understanding the primary effects, and future

research endeavors could systematically explore the introduction of higher-order terms to unravel more intricate relationships within the dataset.

The identification of specific patterns, anomalies, or outliers in the first-order model may guide the selection of variables for introducing higher-order terms, offering a targeted approach to model refinement. This avenue of research becomes particularly compelling when supported by a well thought out theoretical framework that justifies the inclusion of non-linear components. By considering the potential impact of higher-order terms, future research could delve deeper into the complexities of the relationships between variables, enhancing the sophistication of the model without sacrificing its interpretability. Such endeavors would contribute to a more comprehensive understanding of the relationships being studied and further refine and improve the applicability of the model in real-world scenarios. With these thoughts in mind, we begin the validation of the final first order MLR below.

$$\mathbb{E}[MEDV] = \beta_0 + \beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT} + \beta_8 \cdot \text{RAD} + \beta_9 \cdot \text{CHAS}$$

Chapter 6

Model Validation

Building on the foundation laid in the Data Summary and Preprocessing sections, the Model Validation phase converges theoretical precision with real-world applicability, securing the dependability and predictive efficacy of the finalized first order MLR model. In this chapter, the model's performance is critically assessed, the impact of various influential factor is scrutinized, and the model is further refined as needed. This section encompasses five key aspects, each contributing to the validity of our analysis. It begins with a detailed exploration of residual analysis, examining the behavior of residuals to ascertain the model's adherence to the underlying model assumptions. Subsequently, strategies for handling outliers, addressing multicollinearity among predictors, and transforming the response variable when necessary will be investigated. Finally, the section culminates with an evaluation of the model's performance on unseen data, providing insights into its predictive capabilities and generalizability. The comprehensive exploration of these elements not only ensures the model's statistical integrity but also enhances its applicability in real-world scenarios. As residual diagnostics, outlier management, multicollinearity resolution, potential transformations, and performance testing are explored; the main research goal of optimizing a MLR model is being addressed through the continual refinement of the previously established model for a more accurate representation of the nuances and complexities within the

original Boston housing market dataset.

Residual Analysis

Residual analysis serves as the initial lens through which we criticize the performance of the MLR model. Residuals, representing the differences between observed and predicted values, offer a unique perspective on the model's adherence to the key assumptions that allow for a regression to take place. Through diagnostic plots such as scatterplots of residuals against fitted values ($\hat{y}_i$), residuals against predictors, normal probability plots, Q-Q plots, histograms, and various statistical measures; valuable insights can be gained into the distributional properties of the errors produced by the model. This study only evidences the residual against fitted values plot during this residual analysis in order to avoid redundant information. This meticulous examination allows us to identify patterns, outliers, or potential violations of model assumptions that are based on residuals. Recall these assumptions are as follows:

- **Homoscedasticity:** $\text{Var}(\varepsilon_i) = \sigma^2$
- **Expected Value:** $\mathbb{E}[\varepsilon_i] = 0$
- **Independently Identically Distributed:** $\varepsilon_i \sim \text{IID}(\mu = 0, \sigma^2)$
- **Normality:** $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

The Residual vs. Fitted plot is an essential diagnostic tool for evaluating the performance of an regression model. This scatter plot representation of the residuals against the predicted values $\hat{y}_i$, offers a comprehensive view of the model's behavior. A Residual vs. Fitted plot that does not suggest a violation of the model assumptions should exhibit a random scatter of points with no discernible patterns, indicating that the model captures the underlying relationships in the data. The 'vertical' spread of the residuals should remain approximately constant as the plot is scanned from left to right. This signifies that the homoscedasticity assumption is being met ($\text{Var}(\varepsilon_i) = \sigma^2$).

Furthermore, if the points of the plot have a 'vertical' average of 0, we can conclude that the expected value of the residuals is 0 ($\mathbb{E}[\varepsilon_i] = 0$). With this idyllic plot in mind, observe the following Residual vs. Fitted plot generated using R on the training data.

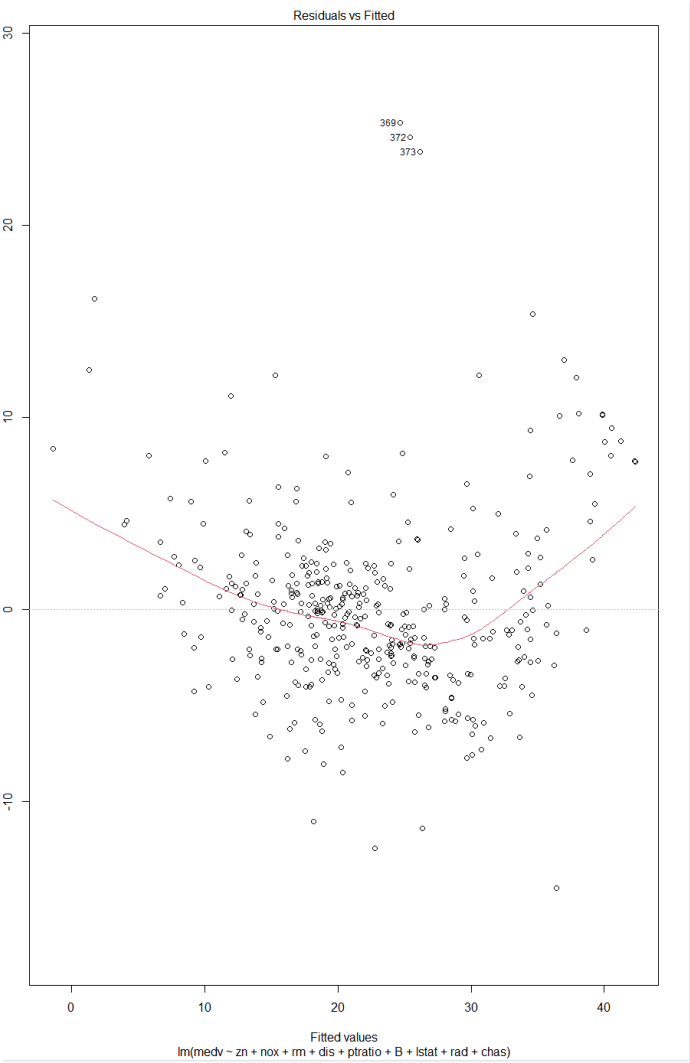

Figure 6.1: The Residual vs. Fitted plot, depicted above, reveals a relatively normal distribution of residuals and maintains a seemingly constant variance. Notably, the presence of outliers is evidenced by the 3 excessively outlying points seen at the top of the plot.

After the model's residuals have been analyzed visually, a couple statistical measures can be used in conjunction with the visual aids to come to a conclusive decision regarding the model assumptions about the residuals. The first test introduced, the Shapiro-Wilk test, is a statistical method employed to

evaluate the normality of the residuals in a MLR model. This test provides a formal assessment of whether the residuals follow a normal distribution. In our analysis, the Shapiro-Wilk test statistic, denoted as W, is calculated to quantify the deviation from normality. The test statistic equation and its related hypothesis statements are as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \epsilon_{(i)}\right)^2}{\sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2}$$

H_0: The residuals are normally distributed

H_a: The residuals are not normally distributed

With this test statistic and its respective null and alternative hypothesis statements defined, observe the following output generated by R.

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.90523, p-value = 2.998e-15
```

Figure 6.2: The Shapiro-Wilk test results in a failed normality test. The calculated test statistic falls below the critical threshold, providing evidence against the assumption of normality in the residuals.

Another statistical measure that can be used to aid the performance of residual analysis is the Durbin-Watson test for autocorrelation and independence. It is employed to assess the presence of autocorrelation or serial correlation in the residuals of a regression model. Autocorrelation occurs when there is a correlation between the residuals at different time points, indicating a violation of the independence assumption ($\varepsilon_i \sim \text{IID}(\mu = 0, \sigma^2)$). Interpreting the Durbin-Watson test involves comparing the calculated d to critical values. If d is close to 2, it supports the null hypothesis of no autocorrelation. If d deviates significantly from 2, it provides evidence for the presence of autocorrelation.The test statistic d is calculated using the following formula which corresponds to the hypothesis tests below:

$$d = \frac{\sum_{j=1}^{n}(\hat{y_j} - \hat{y_{ji}})^2}{(k+1) \cdot MSE}$$

H_0: No autocorrelation among residuals

H_a: Auto correlation exists among residuals

With this test statistic and its respective null and alternative hypothesis statements defined, observe the following output generated by R.

```
> print("Durbin-Watson Test:")
[1] "Durbin-Watson Test:"
> print(dw_test)
 lag Autocorrelation D-W Statistic p-value
   1       0.3978656      1.193869       0
 Alternative hypothesis: rho != 0
```

Figure 6.3: The Durbin-Watson test reveals evidence of autocorrelation as the calculated test statistic falls outside the expected range. Deviations from the ideal value of 2 suggest a failure to meet the assumption of independence in the residuals. This finding prompts a reevaluation of the model and consideration of strategies to address autocorrelation, However, the test statistic is in the 'grey' area as it does not deviate significantly enough from the value of 2 to provide conclusive evidence that the residuals are not independent.

In conclusion, the visual examination of residuals through scatter and diagnostic plots initially suggested an alignment with model assumptions, portraying a seemingly well-behaved random pattern. However, the rigor introduced through statistical tests - such as the Shapiro-Wilk normality test and the Durbin-Watson test for independence - unveiled nuances not immediately apparent from visualizations. The discrepancy between visual impressions and statistical findings underscores the importance of a comprehensive approach to model diagnostics. Notably, the presence of outliers can exert extreme influence on test statistics, potentially leading to conflicting results. Consequently, we pivot our focus to the critical task of handling outliers, recognizing their potential impact on the reliability and validity of our MLR model. By addressing outliers, the aim is to refine the analysis and fortify the model against potential distortion.

Handling Outliers

Now that an in-depth residual analysis has been performed, the focus of Model Validation shifts towards identifying and removing data points which have a significant impact on the model without contributing relevant information to predict the expected value of $MEDV$. Outliers, AKA influential data points that deviate significantly from the general pattern, have the potential to ex-

ert disproportionate impact on statistical analyses. This in turn can distort parameter estimates and compromise the integrity of research findings. Recognizing the significance of outlier management and the seemingly infinite number of ways to handle the presence of outliers, this subsection utilizes 2 very basic strategies aimed at identifying, assessing, and addressing influential observations while also distinguishing them from actual outliers. The overarching objective is to enhance the reliability of the analysis and ensure that the resulting model reflects the genuine relationships inherent in the data, unswayed by the influence of influential data points. With this in mind, an important distinction of the different components that comprise an outlier is necessitated, as the various attributes of a data point are what ultimately decide if it should be accounted for in a given model.

Distinguishing between influence, leverage, and outliers is essential for a deep understanding of the impact of individual data points on regression models. Outliers are data points that deviate significantly from the overall pattern of the data and can disproportionately influence the estimation of model parameters. Leverage, on the other hand, refers to the potential of a data point to influence on the fit of the entire regression model. High leverage points may not necessarily be outliers but can strongly affect a regression line. Influence combines aspects of both outliers and leverage, encompassing the ability of a data point to affect the model parameters and, consequently, the fit of the entire model. While outliers, leverage points, and influential points each have distinct characteristics, recognizing their interplay is crucial for implementing targeted strategies to handle outliers.

With this background understanding of what makes an outlier an outlier, two statistical measures that were used to identify outliers can be introduced.

These measures, known as Cook's distance and leverage, played a crucial role in this evaluation. Cook's distance gauges the impact of each observation on the overall regression model by considering both the magnitude of residuals and the leverage of data points. A high Cook's distance suggests influential observations that can disproportionately affect the model's parameters. Leverage, on the other hand, identifies observations that potentially have a strong pull on the regression line. When used in tandem, Cook's distance and leverage provide a framework for pinpointing observations with a substantial influence on the model, offering a strategic approach to outlier management.

Cook's Distance

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1)\cdot \text{MSE}}$$

Cook's Distance, denoted as D_i, summarizes how much all the values in the regression model change when the i^{th} observation is removed. Bigger values suggest influential points.

- Technically, $D_i > F_{(0.5,k+1,n-(k+1))}$ suggests an influential point

Leverage

$$h_{ii} = x_i(X^T X)^{-1} x_i^T$$

Leverage, represented as h_i, gauges the impact of each observation on its own predicted value. It measures how much a single data point affects the fitted values.

- X: The design matrix containing all predictors.
- Technically, $h_{ii} > \frac{2\cdot(k+1)}{n}$ suggests a potential leverage point

With these measures defined, below are some visual representations of observations and their respective Cook's distances and leverage values.

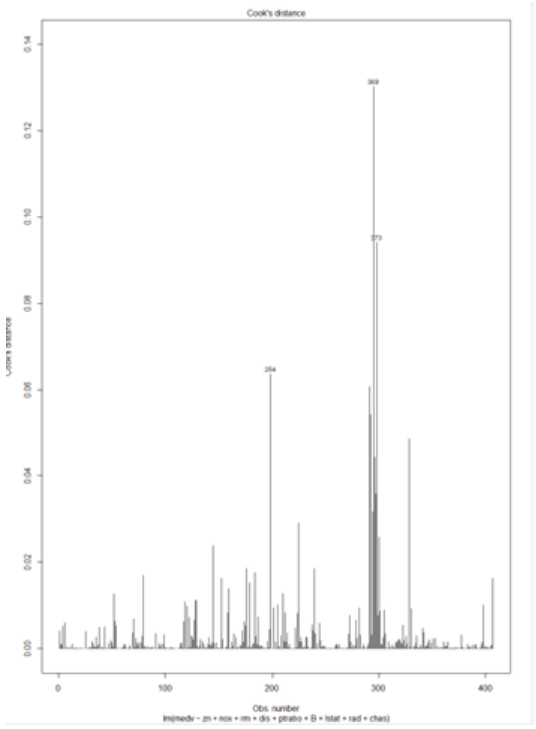

Figure 6.4: The Cook's Distance plot visualizes the influence of each observation on the regression model. Points with larger Cook's Distance values indicate potential outliers or influential data points. While a few observations obviously stand out, the overall presence of outliers appears moderate.

In the pursuit of refining the MLR model, a strategic approach was employed to identify and subsequently remove influential observations from the training data. Using R, a list of observation numbers was generated based on two critical criteria: observations with a leverage value exceeding the predefined threshold and Cook's distance surpassing that respective specified threshold. This dual criterion allowed for a stringent selection of observations with both high leverage and substantial influence on the regression model. Subsequently, the identified observations were systematically removed from the training dataset, mitigating their disproportionate impact on parameter estimation and fostering a more robust and precise regression model. This targeted approach ensures the integrity of the analysis and also reflects a specific strategy for handling influential data points in the context of MLR. The list of observation numbers removed from the training dataset is as follows:

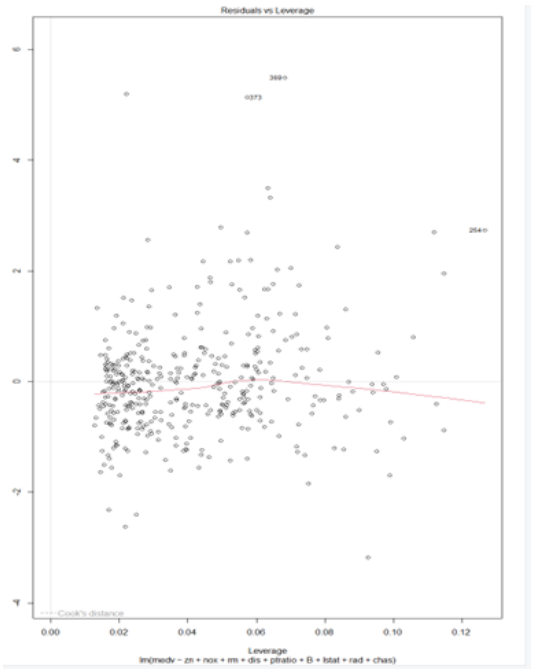

Figure 6.5: The Residuals vs. Leverage plot illustrates the relationship between standardized residuals and leverage, visualizing potential outliers and influential data points. Some observations exhibit higher residuals and leverage, indicating potential outliers. The same 3 outliers can be seen on both the Residuals vs Leverage plot and the Cook's Distance plot.

```
> print(outliers)
[1] "153" "254" "284" "302" "365" "366" "368" "369" "373"
```

Figure 6.6: List of observation numbers which contained outliers according to Cook's Distance and Leverage measures over specified threshold. These were the observation numbers removed from the training data.

The strategic removal of the nine outliers identified above via the outlier handling process outlined in this subsection significantly contributed to the optimization of the MLR model. By systematically addressing observations with disproportionate influence on the regression parameters, we have effectively mitigated the potential distortions introduced by these outliers. Their absence from the training data not only aligns with a more accurate representation of the underlying patterns within the Boston housing market dataset but also lays the groundwork for a more resilient and reliable model. As we transition to the next phase of our analysis, focused on multicollinearity checks, the streamlined dataset free from the undue influence of outliers provides a

solid foundation for an examination of the interrelationships among predictor variables. This approach ensures that the MLR model is unburdened by the impact of influential yet aberrant observations. At this point in research, all of the previous residual analysis steps were performed to ensure that the removal of outliers had a positive impact on the model. Although the residuals were still not perfectly normal, an improvement in all visual diagnostics was evident as well as a Durbin-Watson test statistics closer to 2 and a Shapiro-Wilk normality test that failed to reject the null hypothesis. The resulting R output is not provided in this study to minimize redundant information throughout the thesis. This will become common practice over the next several subsections due to the number of diagnostics used during residual analysis and the need to check residuals after each model refinement. For now, all that matters is that the model is seeing improvement in regard to the underlying assumptions about it's residuals. With this in mind, we begin to check our model to see if predictors are contributing redundant information - multicollinearity.

Checking for MultiCollinearity

The focus of research now shifts towards highlighting the interdependencies among predictor variables within the MLR model. Multicollinearity, the phenomenon where predictor variables are highly correlated, can pose challenges in estimating the individual contributions of each variable - one of the main goals of this research. To systematically assess the degree of intercorrelation, we employ two complementary methods. First, we investigate pairwise correlation coefficients among predictors via a correlation coefficient matrix, seeking to identify instances of strong correlations that may compromise the model's precision. Additionally, we examine the variance inflation of predictor variables, known as the Variance Inflation Factors (VIF), the VIF is a quantitative

measure that illuminates the extent of multicollinearity. These methods collectively empower us to discern intricate relationships among predictors, ensuring that the predictors of the model are not functions of other predictors included in the model.

To begin examining the multicollinearity among the predictor variables in the model, the first assessment used is the pairwise correlations among predictor variables. The cor.test(method = "pearson") function in R is employed, which not only provides correlation coefficients but also associated p-values. The Pearson correlation coefficient, denoted as ρ, ranges from -1 to 1, with 1 indicating a perfect positive linear relationship, -1 indicating a perfect negative linear relationship, and 0 signifying no linear correlation. The corresponding p-values help us determine whether the observed correlation is statistically significant. The null and alternative hypothesis statements that the p-values will determine the validity of are as follows:

$$H_0 : \rho = 0\text{: no significant correlation between the pairs of predictors.}$$
$$H_a : \rho \neq 0\text{: significant correlation between pairs of predictors}$$

Given the hypothesis above, the following r-output provides insight into the ability to reject or accept these statements. Note that the 7 variables shown in the correlation matrix below correspond to the numeric main effects given by $\beta_1 \cdot \text{ZN} + \beta_2 \cdot \text{NOX} + \beta_3 \cdot \text{RM} + \beta_4 \cdot \text{DIS} + \beta_5 \cdot \text{PTRATIO} + \beta_6 \cdot \text{B} + \beta_7 \cdot \text{LSTAT}$. The reason for this being that the Pearson correlation coefficienct, ρ, can not be computed for categorical variables. Further, it doesn't make sense to include the intercept in the matrix as it has high multicollinearity with all predictor variables.

```
> # print matrices of pairwise correlation coefficients and respective p - values
> print(cor_matrix)
           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]
[1,]         NA -0.5131850  0.2824357  0.6662700 -0.3964391  0.1825304 -0.4064165
[2,] -0.5131850         NA -0.3187899 -0.7744653  0.1956096 -0.3799065  0.5926400
[3,]  0.2824357 -0.3187899         NA  0.2209344 -0.3472126  0.1450127 -0.6250343
[4,]  0.6662700 -0.7744653  0.2209344         NA -0.2339807  0.2994708 -0.5021039
[5,] -0.3964391  0.1956096 -0.3472126 -0.2339807         NA -0.2012524  0.3825381
[6,]  0.1825304 -0.3799065  0.1450127  0.2994708 -0.2012524         NA -0.3833043
[7,] -0.4064165  0.5926400 -0.6250343 -0.5021039  0.3825381 -0.3833043         NA
> print(pvalue_matrix)
           [,1]      [,2]      [,3]       [,4]      [,5]       [,6]      [,7]
[1,]         NA 0.4994112 0.1565591 0.08599143 0.3084592 0.14239657 0.1812810
[2,] 0.34471684        NA 0.4091670 0.09946973 0.1122943 0.41099165 0.5439708
[3,] 0.53995859 0.2923747        NA 0.24020855 0.5144807 0.27829537 0.4392257
[4,] 0.08660052 0.6016629 0.6291048         NA 0.5620380 0.42619043 0.4325670
[5,] 0.18092668 0.4007197 0.5556136 0.21459225        NA 0.18606753 0.5913325
[6,] 0.31855723 0.3203985 0.5155291 0.11485707 0.4134675         NA 0.1920514
[7,] 0.43925701 0.1107987 0.1113797 0.58973960 0.1222381 0.09034346        NA
```

Figure 6.7: A correlation matrix where none of the pairwise correlation coefficients have p-values which provide sufficient evidence to reject the null hypothesis statement. Thus, there is no evidence of multicollinearity based on this assessment.

Initial examination of the correlation matrix suggests an absence of strong pairwise correlations among predictor variables. However, to ensure a thorough assessment of multicollinearity, we now turn our attention to the Variance inflation factor (VIF).

VIF is a metric that assesses the degree to which the variance of an estimated regression coefficient is inflated due to multicollinearity. The VIF for each predictor is calculated by regressing that predictor against all other predictors. In general, a larger VIF indicates multicollinearity, where a VIF ≥ 4 is a potential concern which dictates further investigation. A VIF value ≥ 10 indicates a major problem with multicollinearity. This method complements the correlation coefficient analysis. Using R, the VIF of each predictor variable was calculated as follows:

	GVIF	Df	GVIF^(1/(2*Df))
zn	2.391702	1	1.546513
nox	3.856674	1	1.963842
rm	1.897369	1	1.377450
dis	3.724616	1	1.929926
ptratio	2.210354	1	1.486726
B	1.344935	1	1.159714
lstat	2.629286	1	1.621507
rad	4.850445	8	1.103726
chas	1.069453	1	1.034144

Figure 6.8: The VIF plot reveals the multicollinearity assessment for predictor variables in the regression model. Notably, only one predictor exhibits a VIF exceeding the threshold of 4, suggesting a potential concern for intercorrelation. This can be explained however by the fact that the variable with such a VIF is RAD, which obviously has multicollinearity due to it being the only categorical predictor in the model with more than 2 categories. Therefore, there is no significant evidence of variance inflation.

To conclude the assessment of multicollinearity, our scrutiny of correlation coefficients and VIF values across predictor variables has yielded reassuring unanimous results. The correlation matrix indicated no substantial pairwise correlations between predictors, and the VIF analysis confirmed that multicollinearity is not a concern within our regression model. With this confirmation, we pivot our focus to the transformative step of addressing the response variable. The absence of multicollinearity provides a solid foundation for the exploration of the next stage of research, which aims to further enhance the model's predictive power by exploring potential transformations to the response variable.

Transformation of Response Variable

In this section dedicated to transforming the response variable, the research focus narrows to the application of the Box-Cox transformation. As a powerful tool in statistical modeling, the Box-Cox transformation serves to stabilize variances and normalize residuals, aligning the response variable with the assumptions of multilinear regression. Specifically tailored to handle cases where the response variable exhibits non-constant variance and non-normality,

the Box-Cox transformation allows for a systematic exploration of parameter values to achieve optimal data normality. Central to this method is the introduction of a power parameter, denoted as lambda (λ), which governs the transformation applied to the response variable. The transformation is expressed as the following:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

The key challenge lies in determining the optimal value for λ. To this end, the Box-Cox transformation aims to maximize the log-likelihood function, effectively seeking the value that maximizes the normality of the transformed data. The log-likelihood function is given by:

$$\ell(\lambda) = \left(1 - \frac{1}{n}\right) \sum_{i=1}^{n} \log(y_i(\lambda)) - \frac{\lambda}{2} \sum_{i=1}^{n} \left(\frac{\log(y_i(\lambda))}{\log(e)}\right)^2$$

Where n is the sample size, y_i s the individual observation, and e is the base of the natural logarithm. The optimal λ is the value that maximizes this log-likelihood function, which strikes a balance between variance constancy and normality.

In practice, R facilitates this process by automatically calculating the log-likelihood across a range of λ values, allowing for the identification of the optimal transformation. The successful application of the Box-Cox transformation not only yields a more suitable representation of the response variable but also enhances the overall performance and reliability of the MLR model. Now that the concept of a Box-Cox transformation has been introduced, the identification of the optimal λ value that maximizes the log-likelihood function can be outsourced to R. This key λ parameter plays a pivotal role in achieving variance stabilization and normality. The following R output is a result of the boxcox() function on the full finalized model using the training data without

outliers. It should be noted that the same transformation was tested with outliers included in the same data set, and it had no impact whatsoever on the optimal λ value.

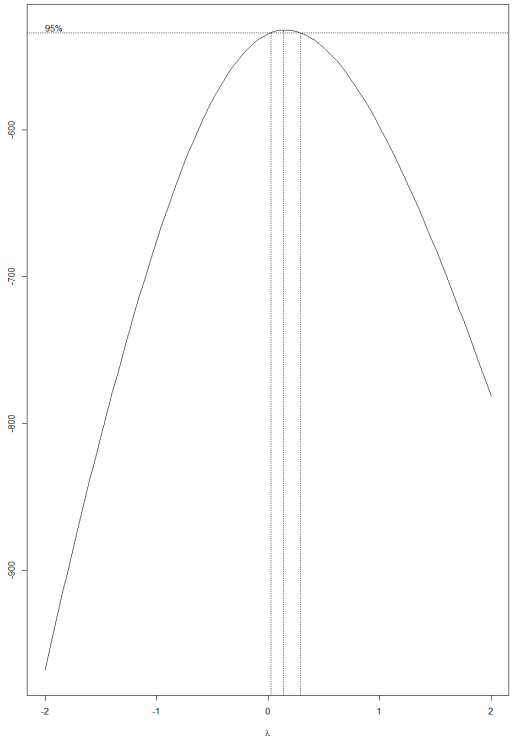

Figure 6.9: Graph showing the optimization curve of different λ values for the given Box-Cox transformation.

```
> lambda <- boxcox_result$x[which.max(boxcox_result$y)]
> print(lambda)
[1] 0.1414141
```

Figure 6.10: The actual optimal value of $\lambda = 0.1\bar{4}$ based on sample data in the training set.

This parameter was applied using the Box-Cox transformation piecewise functions to transform the response variable, further aligning it with the model assumptions of MLR. The tailored adjustment aims to achieve variance stabilization and enhance the normality of the response variable, ultimately contributing to the accuracy and integrity of the regression model. To validate the effectiveness of this transformation, we turn to visual diagnostics. The following histogram and Quantile-Quantile (Q-Q) plot serve as compelling evidence,

offering a comparative view of the distribution of the original and transformed response variable. These graphical representations will illuminate the success of the Box-Cox transformation in rendering the data more amenable to linear modeling assumptions.

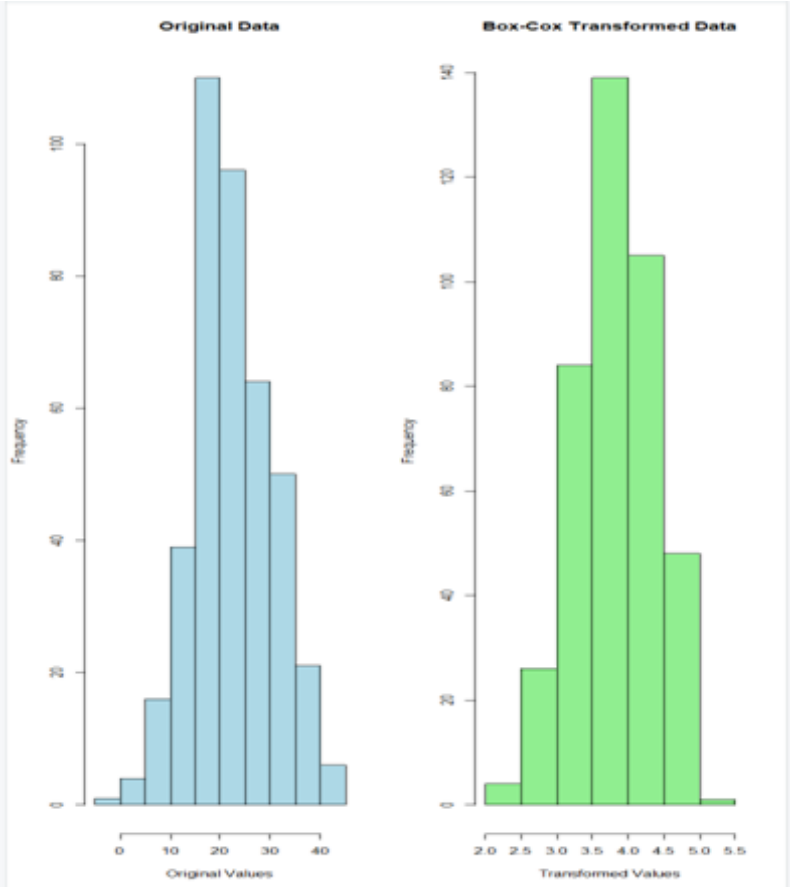

Figure 6.11: Histograms of the original distribution of $MEDV$ (left) vs. the transformed distribution of $MEDV$ (right). Notice that the histogram of the transformed response follows a normal distribution better than the original.

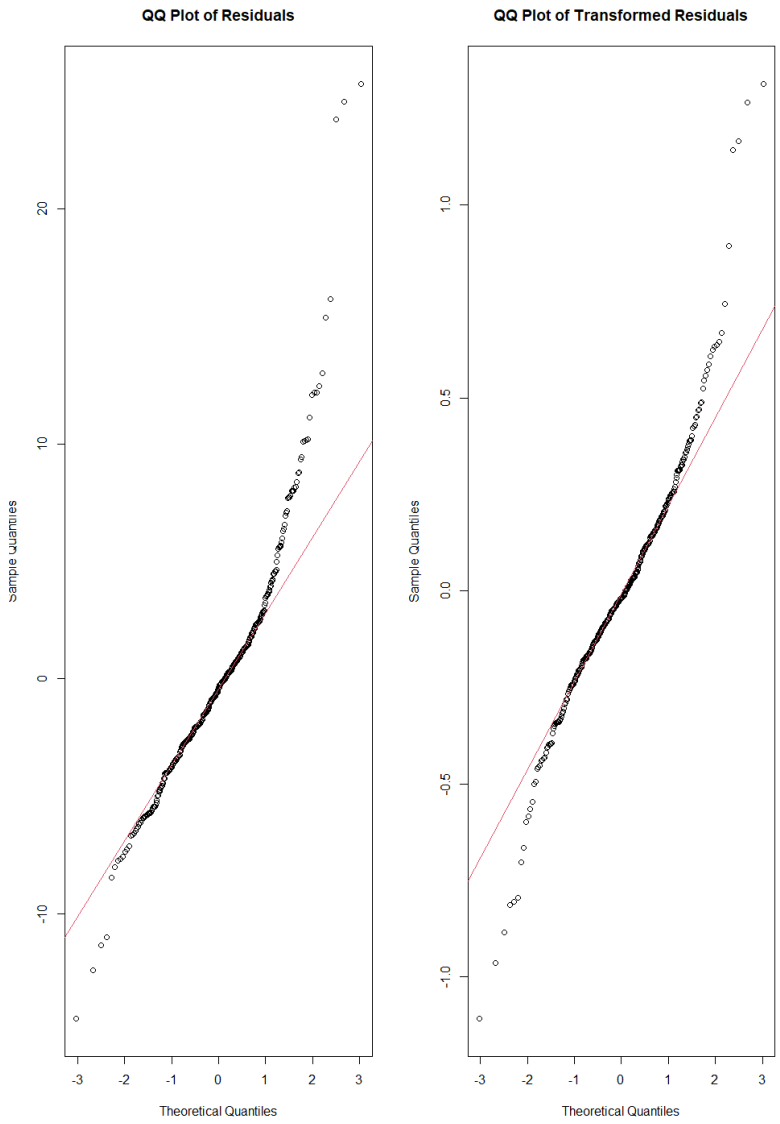

Figure 6.12: Q-Q plots of residuals for $MEDV$ (left) vs. the residuals of the transformed $MEDV$ (right). Notice that the transformed Q-Q plot has more points on the straight line and is more symmetric at the very end when the data points move away from the line. This further validates that the transformation was successful.

As the research transitions from the transformative phase of our analysis, it is evident that the Box-Cox transformation has played a pivotal role in aligning our response variable with the prerequisites of linear regression. The success of this endeavor is notably reflected in the improved normality and variance stabilization achieved. This is evidenced by the R output above. Alluding back to a point brought up during outlier handling, there exist many ways of

validating the improvements of the model brought about by the transformation. Only some are evidenced here to avoid redundancy. In reality an entire residual analysis was performed on the transformed model, which did nothing other than further validate everything seen up to this point in research, hence why it is redundant information.

Armed with a refined and more reliable MLR model, the study focus changes gears towards testing the model's performance on previously unseen data. This crucial step involves not only evaluating the model's predictive accuracy, but iy also necessitates an inverse transformation of the response variable for meaningful interpretation of the results. The inverse transform allows us to seamlessly revert the transformed predictions to their original scale. This approach is used in conjunction on the same tests and visuals as the finalized non-transformed model to final conclusions about the model during validation.

Testing Model on Unseen Data

Now that the MLR model has been refined through the Box-Cox transformation, the next step in is to assess its predictive performance on previously unseen data. The test dataset, which was seperated from the training during preprocessing, serves as a crucial benchmark for evaluating the model's generalization capabilities. Two distinct approaches were employed in making predictions on the test data: utilizing the original response variable and applying the inverse transform to predictions based on the transformed response variable. This dual strategy allows for a comprehensive evaluation of the model's effectiveness both before and after the transformation.

Upon evaluation of prediction accuracy, both Mean Squared Error (MSE) and Root Mean Squared Error ($RMSE$) were calculated for predictions based on the original and transformed response variables. The comparison revealed that, overall, the model's predictions were more accurate when based on the transformed response variable. This outcome aligns with the expectations set by the Box-Cox transformation, emphasizing its role in improving the precision of the regression model. The resulting MSE and $RMSE$ of both models being ran against unseen data can be seen below.

```
Mean Squared Error (MSE): 22.37919
> cat("Root mean Squared Error (RMSE):", rmse, "\n")
Root mean Squared Error (RMSE): 4.730665
> cat("Mean Squared Error (MSE) of transformed model:", MSE, "\n")
Mean Squared Error (MSE) of transformed model: 18.68423
> cat("Root mean Squared Error (RMSE) of transformed model:", RMSE,
"\n")
Root mean Squared Error (RMSE) of transformed model: 4.322526
```

Figure 6.13: MSE and $RMSE$ of both the transformed and original model on test data. Both values are low, indicating that the overall error present in the model is relatively small. Note that the MSE for the transformed model is even lower.

Both MSE and $RMSE$ share a common interpretation — lower values indicate better predictive performance. A model with lower MSE and $RMSE$ values implies that, on average, its predictions are closer to the true values in the dataset. Therefore, in the context of our MLR model, lower MSE and RMSE values indicate that the modeling process was successful

Having quantified the model's predictive accuracy through Mean Squared Error and Root Mean Squared Error, we now turn our attention to a visual inspection of the model's performance. Predicted vs. Actual plots serve as an intuitive representation, offering a comparative view of the model's predictions against the true values in the test dataset. These scatter plots provide a general understanding of how well the model aligns with the actual observations, highlighting potential patterns, trends, or outliers that may not be immediately apparent through numerical metrics alone. The following plots

represent the values predicted by both the original and transformed model. The realm of model testing is expansive and dynamic, offering many avenues

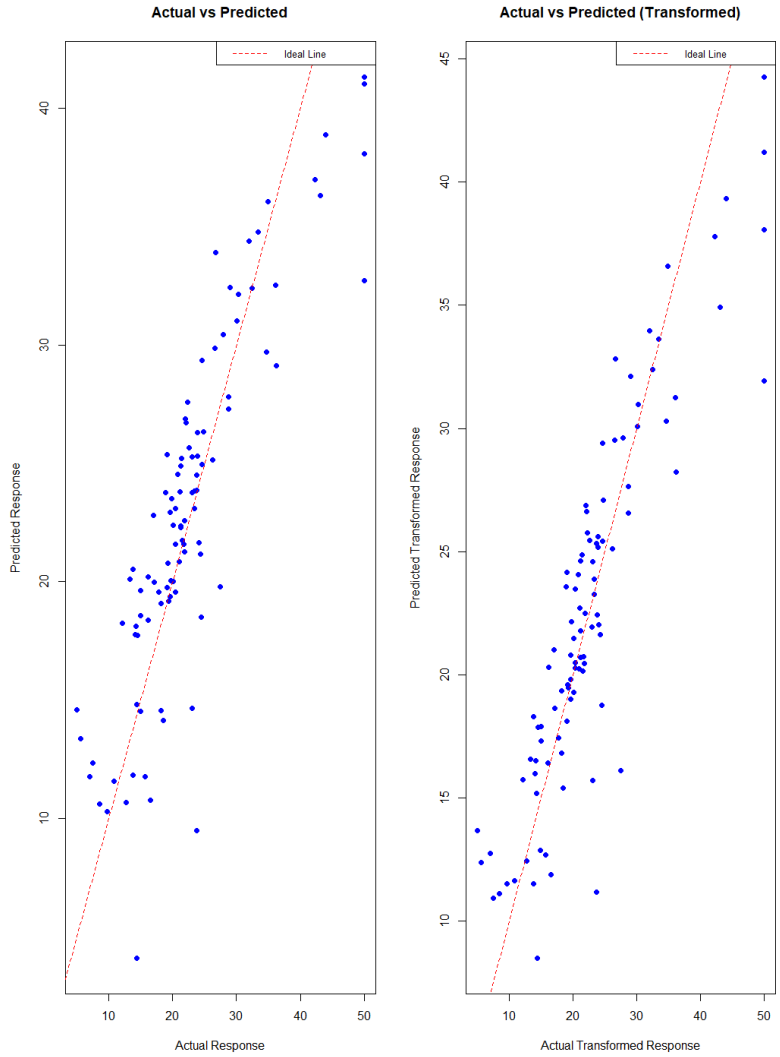

Figure 6.14: The Predicted vs. Actual plot visually encapsulates the performance of the MLR models on the test dataset. Each point represents a distinct observation, with the x-axis denoting the actual values given by the test dataset and the y-axis showcasing the corresponding predictions. A well-behaved scatter of points, closely hugging the diagonal line, signifies the model's accuracy in capturing the underlying relationships within the Boston housing market data. This alignment between predicted and actual values serves as a visual testament to the model's precision. It also further reinforcing the usefulness of the Box-Cox transformation in refining the predictive capabilities of the MLR, as the plot of Predicted vs. Actual for the inverse transformed model is even better than the original.

of refinement in regards to the the predictive capabilities of a model. Beyond the current evaluation, numerous possibilities exist for testing the model's robustness and generalization to diverse scenarios. Varied splits of the origi-

nal dataset other than the generic 80/20 split used for this study, alternative validation techniques such as k-fold cross-validation, and the exploration of additional predictors or higher order terms present themselves as avenues for further investigation. The flexibility of testing methodologies allows for a comprehensive understanding of the model's behavior under different conditions, unveiling its strengths and potential limitations. Visual diagnostics are continuously utilized to leverage the power of graphical representation in an attempt to uncover patterns, outliers, and the model's adherence to assumptions, adding an additional layer of insight into its performance.

As the landscape of model testing continues to grow far beyond the scope of this research, it becomes evident that the model validation process is not a one-size-fits-all endeavor. Rather, it's an ever-changing exploration that demands a tailored approach to address the unique characteristics of the dataset and research objectives. The infinite permutations of testing methodologies underscore the iterative and evolving nature of model validation. As the conclusion of model validation journey begins, it is crucial to reflect on the insights gained, acknowledge the avenues explored, and outline potential directions for future research to further enhance the reliability and applicability of the MLR model.

In summary, our comprehensive journey through model validation has been marked by meticulous examination and refinement, encompassing various facets of the MLR model. From the beginning foundational stages of residual analysis, handling outliers, and addressing multicollinearity, to the transformative application of the Box-Cox method on the response variable, each step has contributed to a nuanced understanding of the model's intricacies. The evaluation of the model's predictive accuracy, which used metrics such as MSE and RMSE alongside the visual inspection of predicted vs. actual plots, solid-

ifies the model's reliability. The successful integration of these methodologies serves as a testament to the thoughtfulness of this regression analysis on the Boston housing market dataset. However, the dynamic nature of model validation invites ongoing exploration, and the collaborative interplay between theoretical statistics and practical application positions our model as a foundation for continuous improvement. To conclude this chapter, the insights gained and the avenues explored pave the way for future research to delve deeper into alternative techniques, diverse splits, and additional features, ensuring the sustained refinement and applicability of our regression model in real-world scenarios.

Chapter 7

Conclusion

Finalized MLR Models

Parameter Estimates of Final Model Post Processing

```
Call:
lm(formula = medv ~ zn + nox + rm + dis + ptratio + B + lstat +
    rad + chas, data = data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7098  -2.5814  -0.3701   1.7904  25.3605

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.664232   6.032988   5.414 1.09e-07 ***
zn            0.039144   0.016424   2.383 0.017647 *
nox         -17.521762   4.116414  -4.257 2.62e-05 ***
rm            3.493757   0.453461   7.705 1.15e-13 ***
dis          -1.419381   0.221578  -6.406 4.41e-10 ***
ptratio      -0.931891   0.165713  -5.624 3.63e-08 ***
B             0.011452   0.003006   3.810 0.000162 ***
lstat        -0.580015   0.054566 -10.630  < 2e-16 ***
rad2          1.636567   1.596902   1.025 0.306089
rad3          4.737691   1.446463   3.275 0.001152 **
rad4          2.024157   1.268100   1.596 0.111270
rad5          2.841969   1.288949   2.205 0.028060 *
rad6         -0.018012   1.542783  -0.012 0.990691
rad7          3.908164   1.837518   2.127 0.034073 *
rad8          4.846972   1.567729   3.092 0.002136 **
rad24         3.973569   1.429901   2.779 0.005724 **
chas1         2.413211   0.956061   2.524 0.012004 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.794 on 381 degrees of freedom
Multiple R-squared:  0.7351,    Adjusted R-squared:  0.724
F-statistic: 66.08 on 16 and 381 DF,  p-value: < 2.2e-16
```

Figure 7.1: summary() output of finalized MLR model. Parameter estimates can be seen under the 'Estimates' column. Each row corresponds to a different parameter estimate, given under the 'Coefficients' column.

```
Call:
lm(formula = ((medv^lambda - 1)/lambda) ~ zn + nox + rm + dis +
    ptratio + B + lstat + rad + chas, data = data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10860 -0.16060 -0.01768  0.14684  1.31506

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.9220039  0.3770994  13.052  < 2e-16 ***
zn           0.0014887  0.0010266   1.450  0.14784
nox         -1.1111136  0.2573016  -4.318 2.01e-05 ***
rm           0.1417901  0.0283442   5.002 8.66e-07 ***
dis         -0.0749941  0.0138500  -5.415 1.09e-07 ***
ptratio     -0.0550211  0.0103581  -5.312 1.85e-07 ***
B            0.0008889  0.0001879   4.731 3.15e-06 ***
lstat       -0.0481500  0.0034107 -14.117  < 2e-16 ***
rad2         0.1610776  0.0998163   1.614  0.10741
rad3         0.2948805  0.0904130   3.261  0.00121 **
rad4         0.1367715  0.0792642   1.726  0.08525 .
rad5         0.1969481  0.0805673   2.445  0.01496 *
rad6         0.0572050  0.0964336   0.593  0.55340
rad7         0.2583626  0.1148563   2.249  0.02505 *
rad8         0.2910717  0.0979928   2.970  0.00316 **
rad24        0.1894782  0.0893777   2.120  0.03465 *
chas1        0.1697684  0.0597598   2.841  0.00474 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2996 on 381 degrees of freedom
Multiple R-squared:  0.7744,    Adjusted R-squared:  0.7649
F-statistic: 81.72 on 16 and 381 DF,  p-value: < 2.2e-16
```

Figure 7.2: summary() output of transformed MLR model. Note the significant increase in the models R_a^2 value.

Full Model

$$\mathbb{E}[MEDV] = 33.664232 + 0.039144 \cdot ZN - 17.521762 \cdot NOX + 3.493757 \cdot RM - 1.419381 \cdot DIS - 0.931891 \cdot \text{ptratio} + 0.011452 \cdot \text{B} - 0.580015 \cdot LSTAT + 1.636567 \cdot RAD_2 + 4.737691 \cdot RAD_3 + 2.024157 \cdot RAD_4 + 2.841969 \cdot RAD_5 - 0.018012 \cdot RAD_6 + 3.908164 \cdot RAD_7 + 4.846972 \cdot RAD_8 + 3.973569 \cdot RAD_{24} + 2.413211 \cdot CHAS_1$$

Transformed Model

$$\frac{\mathbb{E}[MEDV]^{0.141414} - 1}{0.141414} = 4.9220039 + 0.0014887 \cdot ZN - 1.1111136 \cdot NOX + 0.1417901 \cdot RM - 0.0749941 \cdot DIS - 0.0550211 \cdot \text{ptratio} + 0.0008889 \cdot \text{B} - 0.0481500 \cdot LSTAT + 0.1610776 \cdot RAD_2 + 0.2948805 \cdot RAD_3 + 0.1367715 \cdot RAD_4 + 0.1969481 \cdot RAD_5 - 0.0572050 \cdot RAD_6 + 0.2583626 \cdot RAD_7 + 0.2910717 \cdot RAD_8 + 0.1894782 \cdot RAD_{24} + 0.1697684 \cdot CHAS_1$$

Interpretations of Parameter Estimates

$\hat{\beta}_0$: *Intercept* The intercept represents the estimated mean value of the response variable (median value of owner-occupied homes) when all predictor variables are zero. In this context, it does not have a direct practical interpretation since some predictors like room count (RM), distance to employment centers (DIS), and others can't realistically be zero for a given observation.

$\hat{\beta}_1$: ZN For a one-unit increase in the proportion of residential land zoned for large lots over 25,000 square feet, we expect the median value of owner-occupied homes to increase by 39 dollars, holding other variables constant.

$\hat{\beta}_2$: NOX A one-unit increase in nitric oxides concentration (measured in parts per ten million) is associated with an average decrease of approximately 17,520 dollars in the median value of owner-occupied homes, holding other variables constant.

$\hat{\beta}_3$: RM For each additional room in a given house, we expect the median value of owner-occupied homes to increase by 3,494 dollars on average, holding all other variables constant.

$\hat{\beta}_4$: DIS A one unit increase in the weighted distance from one of five Boston employment centers to a given home is associated with a decrease of approximately 1,419 dollars on average holding other variables constant.

$\hat{\beta}_5$: $PTRATIO$ A one-unit increase in the pupil-teacher ratio by town is associated with an average decrease of approximately 932 dollars in the median value of an owner-occupied home in Boston, holding other variables constant.

$\hat{\beta}_6 : B$ For each one-unit increase in the proportion of African American residents, the expected median value of owner-occupied homes increases by 114 dollars on average, holding other variables constant.

$\hat{\beta}_7 : LSTAT$ If the percentage of the population which is lower status increases by 1 percent, an average decrease of approximately 580 dollars in the median value of owner-occupied homes, holding all other variables constant, is expected to occur.

$\hat{\beta}_8 : RAD_2$ The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 2 is approximately 1,637 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta}_9 : RAD_3$ The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 3 is approximately 4,738 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{10}} : RAD_4$ The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 4 is approximately 2,025 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{11}} : RAD_5$ The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 5 is approximately 2,842 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{12}}$: RAD_6 The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 6 is approximately -18 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{13}}$: RAD_7 The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 7 is approximately 3,908 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{14}}$: RAD_8 The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 8 is approximately 4,847 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{15}}$: RAD_{24} The estimated difference in the median value of an owner occupied home between a home with the lowest index of accessibility to radial highways and a home with an accessibility index of 24 is approximately 3,974 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

$\hat{\beta_{16}}$: $CHAS_1$ The estimated difference in the median value of an owner occupied home between a home whose tract does bound the Charles River and a home whose tract does not bound the Charles River is approximately 3,974 dollars on average while holding all other levels of RAD constant and adjusting for other 8 predictor variables.

Further Points of Research

As the understanding of MLR for the Boston housing market dataset advances, future research avenues open up with inviting arms. One key area for exploration is the refinement of outlier handling strategies. Investigating advanced techniques, such as robust regression or machine learning algorithms tailored for outlier detection, could bolster the model's resilience against extreme observations. Moreover, understanding the impact of different outlier-handling approaches on both model performance and interpretability remains a valuable facet of ongoing research.

Moving beyond outlier handling, there lies a very dense field in testing for interactions and higher-order terms. Unraveling potential interactions among predictor variables and assessing the impact of nonlinear relationships could elevate the model's sophistication to a level that cannot be reached by a first order model. Future studies might delve into methods for identifying and testing significant interactions or determining optimal polynomial terms. A continuous enhancing of the ability to capture the intricacies within the data set.

Another point of consideration is alternative methods for model testing. The landscape of future potential research interests widens to consider Bayesian model checking or machine learning model evaluation techniques. These alternatives offer diverse perspectives on model validity and performance as well as providing a comprehensive understanding of its generalization capabilities.

Simultaneously, the exploration of alternative regression models becomes more appealing. Techniques such as decision tree regression, random forests,

or support vector regression offer promise in capturing complex relationships. Understanding how these models perform in comparison to multilinear regression remains a potential avenue for future investigation.

In navigating these succinct yet interwoven research trajectories, a journey to refine and advance regression modeling techniques for the nuanced realities of the Boston housing market dataset must be taken. Each avenue holds promise, collectively contributing to the evolution of predictive modeling in the realm of real estate data.

References

References

1. **Harrison Jr., D., & Rubinfeld, D. L. (1978).** *Hedonic Prices and the Demand for Clean Air.* Journal of Environmental Economics and Management, 5(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

2. **Mendenhall, W., & Sincich, T. (2010).** *A Second Course in Statistics* (7th ed.). Pearson. ISBN: 978-0321598499.

3. **Wackerly, D., Mendenhall, W., & Schaeffer, R. (2014).** *Mathematical Statistics with Applications* (7th ed.). Cengage Learning. ISBN: 978-1285051928.

4. **Kaggle.** *Boston Housing Market Dataset.* Retrieved from https://www.kaggle.com/c/boston-housing

5. **The Pennsylvania State University.** *STAT 462 - Applied Regression Analysis.* Retrieved from https://online.stat.psu.edu/stat462/.

6. **Zach**, (2020). *The Four Assumptions of Linear Regression.* Retrieved from https://www.statology.org/linear-regression-assumptions/

7. **Zach**, (2020). *What is a Partial F-Test?.* Retrieved from https://www.statology.org/partial-f-test/