THOMPSON RIVERS UNIVERSITY

Multivariate Regression Analysis of the Boston Housing Market

Master of Science in Data Science

STAT 5320

(Winter 2024)

By the Group:

Mengyang Zhang - T00696467

Sepideh Mansourigovari - T00732620

Ayisha C O K - T007272585

Seethalakshmy Parakkattu Mani- T00728975

KAMLOOPS, BRITISH COLUMBIA

[March, 2024]

Professor: Dr. Mateen Shaikh

Department of Statistics and Mathematics

Contents

1 Introduction

Urban centers undergoing rapid growth necessitate a profound understanding of housing market intricacies for informed decision-making by policymakers, urban planners, and real estate professionals. The analysis of the Boston-corrected dataset presents a unique opportunity to explore the complex dynamics of urban housing markets in the Boston area. This evolved version of the traditional Boston Housing dataset offers an enriched resource for conducting in-depth statistical analysis and predictive modeling within the realms of machine learning and statistics. Through this project report, we aim to leverage this dataset to unravel the underlying dynamics of the Boston housing market, identify key determinants of housing prices, and shed light on the implications of socio-economic and geographic factors on real estate valuation.

2 Dataset Overview

This study uses Boston-corrected dataset which represents a refined and ethically sensitive evolution version of the original Boston dataset. This corrected version has 506 census tracts and 21 feature variables which contain six more variables such as CMEDV, TOWN, TOWNNO, TRACT, LON, and LAT than the original version, to address mistakes and augment information[Table 1]. This corrected data set offers a comprehensive insight into the demographics, geography, and housing characteristics of the Boston area, which are shown in table 1. Each of these additional variables has various purposes. The LON and LAT variables enable geospatial analysis for location-specific housing trends, while CMEDV provides a corrected median value of owner-occupied homes, overcoming the censoring limitations of the original dataset. While TOWN, TOWNNO, and TRACT allow granular analysis at the town level, facilitating a deeper understanding of local market dynamics and spatial patterns in housing prices across Boston's diverse neighborhoods. This corrected version serves as a benchmark for different areas such as statistical or machine learning tasks, particularly ones which focus on predicting housing prices based on multiple factors, making it a valuable resource for understanding urban housing market influences.

2.1 Preprocessing Data

To implement the methodology within the context of the provided dataset, data preprocessing is necessary. This entails the removal of five variables: OBS, MEDV, B, LSTAT, and TOWN. OBS is removed as it

serves as an index variable. MEDV, representing an earlier iteration of CMEDV, is excluded. TOWN, being a categorical variable, is also eliminated. Additionally, the variables B and LSTAT are omitted due to ethical considerations. Further details regarding the removal of B and LSTAT will be expounded upon in the subsequent paragraph.

2.2 Ethical Issues Regarding Variables B and LSTAT:

As previously indicated, the decision to exclude variables B and LSTAT stems from profound ethical considerations pertaining to these attributes. These variables, intended to capture systemic racism and class disparities, encapsulate societal biases that may perpetuate inequities and discrimination within predictive modeling endeavors. The inclusion of mathematically encoded racial biases within variable B and socio economic inequalities represented by LSTAT underscores significant ethical apprehensions regarding the potential propagation of bias and discrimination in housing price predictions. Recognizing the ethical ramifications associated with the utilization of such variables underscores the imperative of conducting research and modeling practices that prioritize principles of fairness, transparency, and social responsibility. Through the removal of variables B and LSTAT from the dataset, we endeavor to uphold ethical standards and advocate for a more impartial and equitable approach to scrutinizing housing market dynamics within the Boston region.

3 Methodology

Our approach involves the following key steps:

3.1 Exploratory Data Analysis (EDA)

The goal of Exploratory Data Analysis (EDA)[Table 2] is to understand the data by examining its central tendencies (mean, median, mode) and distributions (range, IQR, variance, SD). Outliers can be identified through visual tools like box plots or statistical methods such as z-scores and IQR. Only when there is sufficient evidence that outliers are recorded incorrectly (data is unlikely to happen in real life) should they be removed; otherwise, it's important to preserve the characteristics of the data. Additionally, EDA addresses issues related to missing data, often remedied through imputation techniques to ensure the integrity of the dataset for accurate analysis.

3.2 Variable Selection and Model Optimization

Stepwise Regression

Implement stepwise regression to iteratively add or remove variables based on lowering the BIC (Bayesian Information Criterion). This approach ensures that only relevant predictors are included in the final model.

$$BIC = -2\log(\text{likelihood}) + p\log(n) \tag{1}$$

Where p is the number of parameters and n is the number of observations.

Lasso Regularization

Prevent overfitting by penalizing the absolute size of the coefficients. It adds a penalty term to the loss function, encouraging small coefficient values and feature selection.

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{2}$$

Where $\sum_{j=1}^{p} |\beta_j|$ is the penalty term that encourages sparsity in the coefficient estimates, and λ is the regularization parameter that controls the strength of the penalty term.

3.3 Multiple Regression Analysis:

In this study, a regression model is employed to elucidate the impact of several factors, including square footage, number of bedrooms, and location on house prices within the Boston area. Within this model, the price of a house (CMEDV) serves as the outcome variable, while the aforementioned factors, function as predictors. Notably, one of these predictors, proximity to the Charles River (CHAS), is categorical in nature. This analytical approach facilitates a quantitative assessment of the diverse elements influencing the housing market in Boston. By systematically evaluating the relationships between predictor variables such as square footage, number of bedrooms, and the categorical variable indicating proximity to the Charles River, the regression model aims to provide insights into the dynamics driving housing prices within the Boston area.

3.4 Residual analysis and Assumption Testing:

Residual analysis is performed to evaluate regression assumptions, including linearity, homoscedasticity, independence, and normality. The model is adjusted accordingly to meet these assumptions, enhancing the robustness and reliability of inference. This iterative process ensures adherence to statistical principles, promoting accurate interpretation of regression results. By addressing potential violations of assumptions, the validity of the model is strengthened. Through systematic assessment and adjustment, the regression analysis yields more reliable and interpretable outcomes.

3.5 Model Inference and Validation:

Classical and resampling methodologies are employed to estimate confidence intervals and conduct hypothesis testing on the regression coefficients. Subsequently, the interpretation of individual predictors' effects on housing prices is conducted to evaluate the model validation, with a focus on delineating the model's explanatory efficacy. By conducting thorough statistical analysis, we clarify the importance of predictor variables in understanding the dynamics of housing prices. This process reinforces the predictive and explanatory prowess of the model. This analytical endeavor underscores the validation process by discerning the extent to which predictors contribute to the variability in housing prices, thereby facilitating robust interpretations of the regression model outcomes within the domain of the housing market.

4 Results

- Outliers identified through Interquartile Range (IQR) and boxplot[Table 2]analysis were found to fall within a reasonable range, suggesting accurate recording without evidence of inaccuracies. Retaining these outliers was recommended to preserve valuable data insights and variability within the dataset, with no missing values[Table 3] requiring imputation to maintain dataset integrity during exploratory data analysis (EDA).

- The correlation matrix heatmap[Table 1] revealed a strong positive correlation of 0.9 between "RAD" and "TAX" suggesting a simultaneous increase in both variables. High multicollinearity was noted, which can complicate individual predictor impact determination in multivariate regression, potentially leading to unreliable statistical inferences.

- In the updated stepwise regression output, significant predictors of ”CMEDV” included ”RM”, ”TAX”, ”PTRATIO”, ”CRIM”, ”NOX”, ”DIS”, ”AGE”, ”CHAS”, ”RAD”, and ”ZN,” each shows varying positive and negative associations. However, ”LON” was deemed statistically insignificant with a p-value of 0.79, warranting consideration for exclusion from the model. The stepwise regression automatically eliminates the variable ”TRACT” since the backward and forward selection are done at the same time.

- In the process of backward selection in stepwise regression, ”ZN” was automatically excluded by the model, leading to a decrease in BIC, indicating an enhancement in the model's performance after its removal[Table 4].

- Lasso regression with the regularization parameter 'lambda.lse' [Figure 7]aimed to simplify the model by selecting the largest lambda value within one standard error of the minimum cross-validation error. This approach reduces model complexity, potentially sacrificing some predictive accuracy while enhancing generalization ability and addressing issues of high linearity among variables[Table 6].

- Variables excluded from the Lasso regression model (with coefficients zeroed out) were: TOWN, TRACT, LAT, ZN, INDUS, DIS, and RAD[Table 6] where as variables included in the model (with non-zero coefficients) were: Intercept, LON, CRIM, CHAS, NOX, RM, AGE, TAX, and PTRATIO.

- After conducting Model 1[Table 9], it was observed that two variables were deemed insignificant. Subsequently, an attempt was made to eliminate these variables in Model 2[Table 10], leading to a slight decrease in both R-squared and adjusted R-squared values from 0.6342 to 0.6321 and 0.6284 to 0.6277, respectively. When the predictive power remains relatively consistent, opting for a model with fewer variables is advantageous as it reduces complexity and enhances simplicity

- The residual analysis[Figure 11,12,13], depicts a predominantly linear relationship, with a few potential outliers deviating from the pattern. The Q-Q Plot suggests a normal distribution, although some points at the distribution edges hint at potential outliers or heavy tails. The Scale-Location Plot indicates relatively constant residual variance, with fluctuations along fitted values suggesting possible heteroscedasticity.

- Regarding model inference, the variable ”RM” emerges as a significant positive predictor influencing median home value, while ”NOX” exerts a negative impact. Confidence intervals around coefficients

offer a 95% confidence range for the true coefficient values. The mean test error, approximately 46.82, signifies the average deviation of model predictions from actual values. Cross-validation results emphasize the need for balance model complexity and predictive accuracy[Figure 8], showcasing variable prediction errors across different model complexities.

The multiple regression model for the Boston housing dataset is given by:

$$\begin{aligned}\text{CMEDV} = &-654.40 - 9.26 \times \text{LON} - 0.118 \times \text{CRIM} + 3.535 \times \text{CHAS} \\ &- 9.169 \times \text{NOX} + 6.679 \times \text{RM} - 0.0136 \times \text{AGE} \\ &- 0.00342 \times \text{TAX} - 0.824 \times \text{PTRATIO}\end{aligned}$$

5 Discussions

Exploring non-linear relationships is essential in data analysis, particularly when a variable such as 'age' lacks a significant linear correlation with the outcome. Even if 'age' is not strongly related to the outcome and its removal does not substantially impact the Adjusted R-squared value [Table 9], investigating potential non-linear effects is important for a comprehensive analysis. The impact of age on the dependent variable may not be direct or proportional, potentially leading to changing or more pronounced effects as age increases, indicating a non-linear pattern. In such cases, the inclusion of higher-order polynomial terms for age or the utilization of models capable of capturing complex relationships, such as generalized additive models (GAMs) or spline regression, becomes essential. Visualizing and testing non-linear effects through methods like scatterplots with smooth lines or statistical tests for non-linearity can provide valuable insights into the true nature of the relationship between age and the outcome variable.

References

[1] Boston Housing Dataset: **boston corrected**

A Appendix

Table 1: Description of the Boston Housing dataset variables

Variable	Description
TOWN	Town names (categorical)
TOWNNO	Numeric vector corresponding to TOWN
TRACT	Tract ID numbers
LON	Tract point longitudes in decimal degrees
LAT	Tract point latitudes in decimal degrees
MEDV	Median values of owner-occupied housing (USD 1000)
CMEDV	Corrected median values of owner-occupied housing (USD 1000)
CRIM	Per capita crime
ZN	Proportions of residential land zoned (over 25,000 sq. ft.)
INDUS	Proportions of non-retail business acres
CHAS	1 if tract borders Charles River; 0 otherwise
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average numbers of rooms per dwelling
AGE	Proportions of units built prior to 1940
DIS	Weighted distances to five employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate (per USD 10,000)
PTRATIO	Pupil-teacher ratios
B	1000*(Bk - 0.63)^2($Bk = proportionofblacks$)
LSTAT	Percentage of lower status population

Table 2: Summary Statistics of the Dataset

Statistic	TOWN	TRACT	LON	LAT	CMEDV	CRIM
Min.	0.00	1	-71.29	42.03	5.00	0.00632
1st Qu.	26.25	1303	-71.09	42.18	17.02	0.08205
Median	42.00	3394	-71.05	42.22	21.20	0.25651
Mean	47.53	2700	-71.06	42.22	22.53	3.61352
3rd Qu.	78.00	3740	-71.02	42.25	25.00	3.67708
Max.	91.00	5082	-70.81	42.38	50.00	88.97620
Statistic	ZN	INDUS	CHAS	NOX	RM	AGE
Min.	0.00	0.46	0:471	0.3850	3.561	2.90
1st Qu.	0.00	5.19	1: 35	0.4490	5.886	45.02
Median	0.00	9.69		0.5380	6.208	77.50
Mean	11.36	11.14		0.5547	6.285	68.57
3rd Qu.	12.50	18.10		0.6240	6.623	94.08
Max.	100.00	27.74		0.8710	8.780	100.00
Statistic	DIS	RAD	TAX	PTRATIO		
Min.	1.130	1.000	187.0	12.60		
1st Qu.	2.100	4.000	279.0	17.40		
Median	3.207	5.000	330.0	19.05		
Mean	3.795	9.549	408.2	18.46		
3rd Qu.	5.188	24.000	666.0	20.20		
Max.	12.127	24.000	711.0	22.00		

Table 3: Summary Of Missing Variables

TOWN	TRACT	LON	LAT	CMEDV	CRIM	ZN	INDUS	CHAS
0	0	0	0	0	0	0	0	0
NOX	RM	AGE	DIS	RAD	TAX	PTRATIO		
0	0	0	0	0	0	0		

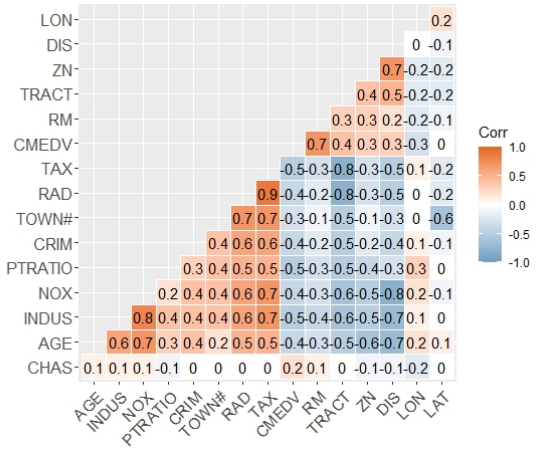


Figure 1: Correlation Between variables

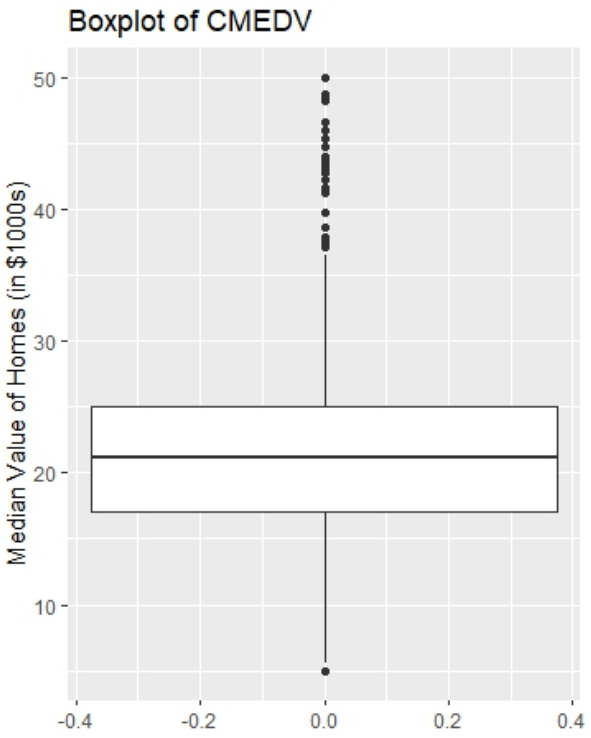


Figure 2: Response Variable- CMEDV

```
16 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept) 22.69798001
TOWN#        .
TRACT        .
LON         -0.13612109
LAT          .
CRIM        -0.63667816
ZN           .
INDUS        .
CHAS         0.55855369
NOX         -0.93492902
RM           4.54776649
AGE         -0.08124037
DIS          .
RAD          .
TAX         -0.25238204
PTRATIO     -1.28149837
```

Figure 3: LASSO Model Summary

```
Call:
lm(formula = CMEDV ~ RM + PTRATIO + NOX + DIS + CHAS + AGE +
    CRIM + TRACT + TAX + RAD, data = boston_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5679 -0.3061 -0.0663  0.2122  4.1641

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.004178   0.029610   0.141  0.88787
RM           0.541167   0.036050  15.012  < 2e-16 ***
PTRATIO     -0.269376   0.038753  -6.951 1.52e-11 ***
NOX         -0.278674   0.060485  -4.607 5.52e-06 ***
DIS         -0.289518   0.051433  -5.629 3.45e-08 ***
CHAS         0.127417   0.031075   4.100 5.02e-05 ***
AGE         -0.129142   0.047519  -2.718  0.00687 **
CRIM        -0.157494   0.059463  -2.649  0.00841 **
TRACT       -0.085933   0.058574  -1.467  0.14315
TAX         -0.190516   0.082246  -2.316  0.02105 *
RAD          0.138805   0.092236   1.505  0.13316
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5886 on 393 degrees of freedom
Multiple R-squared:  0.6598,    Adjusted R-squared:  0.6511
F-statistic: 76.22 on 10 and 393 DF,  p-value: < 2.2e-16
```

Figure 4: Stepwise Model Summary

```
Start:  AIC=-373.35
CMEDV ~ RM + PTRATIO + NOX + DIS + CHAS + AGE + CRIM + TRACT +
    TAX + RAD

          Df Sum of Sq    RSS     AIC
- TRACT    1     0.746 136.91 -377.15
- RAD      1     0.785 136.95 -377.03
- TAX      1     1.859 138.02 -373.88
<none>                 136.16 -373.35
- CRIM     1     2.431 138.60 -372.21
- AGE      1     2.559 138.72 -371.83
- CHAS     1     5.825 141.99 -362.43
- NOX      1     7.355 143.52 -358.10
- DIS      1    10.978 147.14 -348.03
- PTRATIO  1    16.741 152.91 -332.51
- RM       1    78.079 214.24 -196.24

Step:  AIC=-377.15
CMEDV ~ RM + PTRATIO + NOX + DIS + CHAS + AGE + CRIM + TAX +
    RAD

          Df Sum of Sq    RSS     AIC
- TAX      1     1.746 138.66 -378.03
- RAD      1     1.747 138.66 -378.03
<none>                 136.91 -377.15
- AGE      1     2.300 139.21 -376.42
- CRIM     1     2.362 139.27 -376.24
- CHAS     1     5.706 142.62 -366.65
- NOX      1     7.345 144.26 -362.04
- DIS      1    11.150 148.06 -351.52
- PTRATIO  1    15.995 152.91 -338.51
- RM       1    77.473 214.38 -201.98

Step:  AIC=-378.03
CMEDV ~ RM + PTRATIO + NOX + DIS + CHAS + AGE + CRIM + RAD

          Df Sum of Sq    RSS     AIC
- RAD      1     0.216 138.87 -383.40
<none>                 138.66 -378.03
- AGE      1     2.361 141.02 -377.21
- CRIM     1     2.366 141.02 -377.20
- CHAS     1     6.308 144.96 -366.06
- NOX      1     9.309 147.97 -357.78
- DIS      1    11.201 149.86 -352.65
- PTRATIO  1    16.471 155.13 -338.69
- RM       1    80.812 219.47 -198.51

Step:  AIC=-383.4
CMEDV ~ RM + PTRATIO + NOX + DIS + CHAS + AGE + CRIM

          Df Sum of Sq    RSS     AIC
<none>                 138.87 -383.40
- CRIM     1     2.436 141.31 -382.38
- AGE      1     2.507 141.38 -382.18
- CHAS     1     6.398 145.27 -371.21
- NOX      1     9.887 148.76 -361.62
- DIS      1    11.067 149.94 -358.43
- PTRATIO  1    18.499 157.37 -338.88
- RM       1    87.774 226.65 -191.51
```

Figure 5: BIC Model Summary

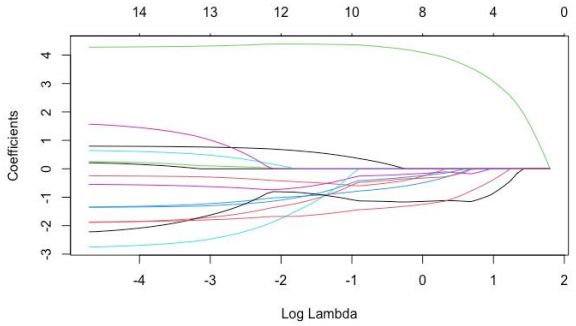


Figure 6: Lasso Model Plot

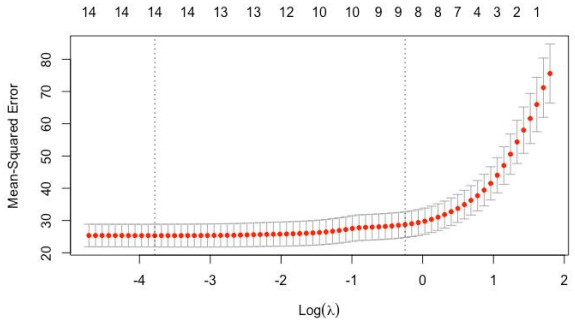
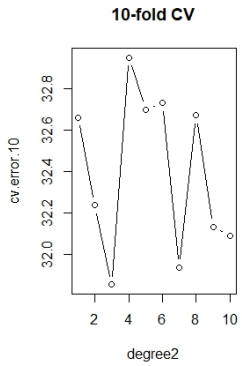


Figure 7: Lasso Model Lambda



(a) K-fold Plot

	2.5 %	97.5 %
(Intercept)	-1.174847e+03	-1.339188e+02
LON	-1.656344e+01	-1.950301e+00
CRIM	-1.886046e-01	-4.769971e-02
CHAS	1.529820e+00	5.540167e+00
NOX	-1.683615e+01	-1.502015e+00
RM	5.897282e+00	7.460952e+00
AGE	-3.979036e-02	1.255731e-02
TAX	-8.361799e-03	1.517863e-03
PTRATIO	-1.114989e+00	-5.348925e-01

Mean test error is: 46.82852

(b) K-fold Summary

Figure 8: K-fold Cross-Validation Plots

```
Call:
lm(formula = CMEDV ~ LON + CRIM + CHAS + NOX + RM + AGE + TAX +
    PTRATIO, data = Boston_houseprice_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-16.613  -2.960  -0.654   1.652  40.174

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.544e+02  2.649e+02  -2.470 0.013835 *
LON         -9.257e+00  3.719e+00  -2.489 0.013130 *
CRIM        -1.182e-01  3.586e-02  -3.295 0.001055 **
CHAS         3.535e+00  1.021e+00   3.464 0.000579 ***
NOX         -9.169e+00  3.902e+00  -2.350 0.019181 *
RM           6.679e+00  3.979e-01  16.785  < 2e-16 ***
AGE         -1.362e-02  1.332e-02  -1.022 0.307217
TAX         -3.422e-03  2.514e-03  -1.361 0.174118
PTRATIO     -8.249e-01  1.476e-01  -5.588 3.79e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.598 on 497 degrees of freedom
Multiple R-squared:  0.6342,    Adjusted R-squared:  0.6283
F-statistic: 107.7 on 8 and 497 DF,  p-value: < 2.2e-16
```

Figure 9: Multiple Linear Regression Model

```
Call:
lm(formula = CMEDV ~ LON + CRIM + CHAS + NOX + RM + PTRATIO,
    data = Boston_houseprice_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-16.702  -3.053  -0.474   1.862  39.418

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -589.62047  255.84481  -2.305 0.021599 *
LON           -8.38169    3.59464  -2.332 0.020112 *
CRIM          -0.13820    0.03314  -4.171 3.58e-05 ***
CHAS           3.58977    1.01460   3.538 0.000441 ***
NOX          -14.04907    2.48811  -5.646 2.75e-08 ***
RM             6.65702    0.39788  16.731  < 2e-16 ***
PTRATIO       -0.93287    0.13101  -7.121 3.77e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.603 on 499 degrees of freedom
Multiple R-squared:  0.6321,    Adjusted R-squared:  0.6277
F-statistic: 142.9 on 6 and 499 DF,  p-value: < 2.2e-16
```

Figure 10: Multiple Linear Regression Model (AGE)

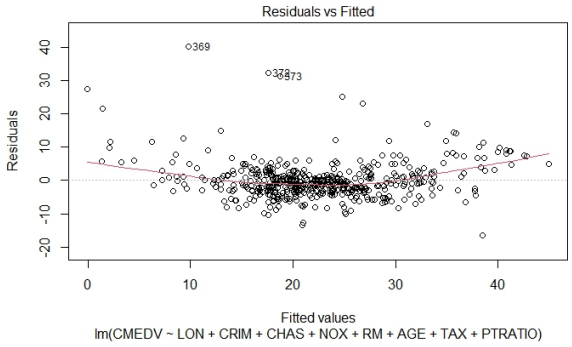


Figure 11: Residuals Vs Fitted

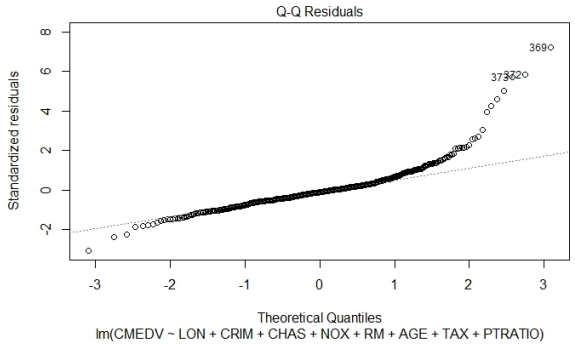


Figure 12: Q-Q Plot

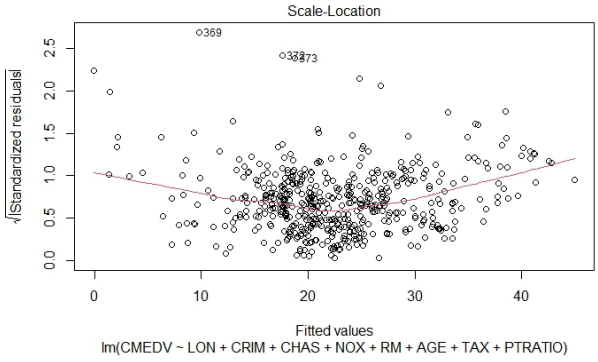


Figure 13: Scale Location

Multiple Linear Regression Analysis: R Code

```r
pkg_list <- c("glmnet", "boot", "MASS", "ggplot2", "openxlsx", "rlang", "readxl",
"dplyr", "cvTools", "Matrix", "boot","caret")
# Install packages if needed
for (pkg in pkg_list) {
  # Try loading the library.
  if (!library(pkg, logical.return = TRUE, character.only = TRUE)) {
    # If the library cannot be loaded, install it; then load.
    install.packages(pkg)
    library(pkg, character.only = TRUE)
  }
}



#(i)  EDA

Boston_houseprice <- read_excel("Boston_houseprice.xlsx")
View(Boston_houseprice)

dim(Boston_houseprice)
names(Boston_houseprice)

sum(is.na(Boston_houseprice))


Boston_houseprice_cleaned <- select(Boston_houseprice, -c(OBS., MEDV, B, LSTAT, TOWN))




# Generate summary for the dataset
boston_summary <- summary(Boston_houseprice_cleaned)
```

```r

# View the summary table
print(boston_summary)


# Identify missing values
missing_values <- sapply(Boston_houseprice_cleaned, function(x) sum(is.na(x)))
missing_values

# Assuming Boston_houseprice is your dataset and CMEDV is the response variable
p <- ggplot(Boston_houseprice, aes(y = CMEDV)) +
  geom_boxplot() +
  labs(title = "Boxplot of CMEDV", y = "Median Value of Homes (in $1000s)")
print(p)
# Statistical method for identifying outliers using IQR
# Identify numeric columns
numeric_columns <- sapply(Boston_houseprice_cleaned, is.numeric)

# Calculate IQR for each numeric column, excluding NA values
IQR_values <- apply(Boston_houseprice_cleaned[, numeric_columns], 2, IQR, na.rm = TRUE)
IQR_values


# Detecting outliers
outliers <- lapply(Boston_houseprice_cleaned, function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR_x <- IQR(x, na.rm = TRUE)
  return(x[x < (Q1 - 1.5 * IQR_x) | x > (Q3 + 1.5 * IQR_x)])
})


# (ii) Variable Selection


Boston_houseprice_scaled <- scale(Boston_houseprice_cleaned)
```

```r
#random sampling
index <- sample(nrow(Boston_houseprice_scaled),nrow(Boston_houseprice_scaled)*0.80)
boston_train <- Boston_houseprice_scaled[index,]
boston_test <- Boston_houseprice_scaled[-index,]

library(ggcorrplot)
corr <- round(cor(boston_train), 1)
ggcorrplot(corr, hc.order = TRUE, type = "lower", lab = TRUE,
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"))

boston_train <- as.data.frame(boston_train)
# Forward variable selection
Model_null <- lm(CMEDV~1, data = boston_train)
Model_full <- lm(CMEDV~., data = boston_train)

Model_step<- step(Model_null, scope=list(lower=Model_null,
upper=Model_full), direction='both', trace=FALSE)
summary(Model_step)


Model_step.BIC <- step(Model_step, k=log(nrow(boston_train)))


set.seed(1234)
# Standardizing the data excluding the target variable 'CMEDV'
Boston_standardised <- scale(dplyr::select(Boston_houseprice_cleaned, -CMEDV))
# Creating training and testing sets
X_train <- as.matrix(Boston_standardised)[index, ]
X_test  <- as.matrix(Boston_standardised)[-index, ]
Y_train <- Boston_houseprice_cleaned$CMEDV[index]
Y_test  <- Boston_houseprice_cleaned$CMEDV[-index]

# Finding the optimal lambda with 10-fold cross-validation

```

```r
cv.lasso <- cv.glmnet(x = X_train, y = Y_train, alpha = 1, family = "gaussian")

# Plotting the cross-validation curve to find the optimal lambda
plot(cv.lasso)


#fit model
Model_lasso<- glmnet(x=X_train, y=Y_train, family = "gaussian", alpha = 1)
plot(Model_lasso, xvar = "lambda")


Cross_validation_lasso<- cv.glmnet(x=X_train, y=Y_train,
family = "gaussian", alpha = 1, nfolds = 10)
plot(Cross_validation_lasso)

lamda_min<-Cross_validation_lasso$lambda.min
lamda_lse<-Cross_validation_lasso$lambda.1se

coef(Model_lasso, s=lamda_min)

coef(Model_lasso, s=lamda_lse)

par(mfrow=c(2,2))
plot(Model_step)

par(mfrow=c(2,2))
plot(Model_lasso)

#(iii) Multiple regression analysis
#Model1
Model_multiple <- lm(CMEDV ~ LON+CRIM+CHAS+NOX+RM+AGE+TAX+PTRATIO,
data = Boston_houseprice_cleaned)
summary(Model_multiple)
#MODEL2
Model_multiple_final <- lm(CMEDV ~ LON+CRIM+CHAS+NOX+RM+PTRATIO,
data = Boston_houseprice_cleaned)
```

```r
summary(Model_multiple_final)

(iv)
plot(Model_multiple)

(v)

# Estimate confidence intervals
conf_int <- confint(Model_multiple, level=0.95)
print(conf_int)


#stepwise_model
model_vars <- names(coef(Model_multiple)[-1])
set.seed(123)
k <- 5 #
n <- nrow(Boston_houseprice_cleaned)
folds <- cut(seq(1,n), breaks=k, labels=FALSE)

Boston_houseprice_cleaned$CHAS <- as.factor(Boston_houseprice_cleaned$CHAS)

test_errors <- vector('numeric', k)

for(i in 1:k){

  testIndexes <- which(folds==i, arr.ind=TRUE)
  testData <- Boston_houseprice_cleaned[testIndexes, ]
  trainData <- Boston_houseprice_cleaned[-testIndexes, ]

  # stepwise_model
  formula_str <- paste("CMEDV ~", paste(model_vars, collapse=" + "))
  model_formula <- as.formula(formula_str)
  model <- lm(model_formula, data=trainData)


  predictions <- predict(model, newdata=testData)
```

```
  test_errors[i] <- mean((predictions - testData$CMEDV)^2)
}


mean_test_error <- mean(test_errors)
cat("Mean test error is:", mean_test_error, "\n")



# k-Fold Cross-Validation
set.seed(17)
cv.error.10=rep(0,10) # 10 degree polynomial

for (i in 1:10){
 glm.fit=glm(Model_multiple,data=Boston_houseprice_cleaned)
 cv.error.10[i]=cv.glm(Boston_houseprice_cleaned,glm.fit,K=10)$delta[1]
 }
cv.error.10

# Plot the CV errors
par(mfrow=c(1,2))
#degree1=1:10
#plot(degree1, cv.error, type="b", main="LOOCV")

degree2=1:10
plot(degree2, cv.error.10, type="b", main="10-fold CV")
```