Boston Housing Data Analysis Report

ZERONG WANG, HAOTING ZHU, CHUQIAO SONG, ZHENJIE CHEN

December 23, 2023

Contents

1 Background

We conduct a regression analysis of the BostonHousing dataset and draw conclusions from the perspective of the government regulating house prices in real time and maintaining house price stability while securing real estate tax revenues.

Land finance in China plays an important role in the economy, with far-reaching impacts on national finances, urbanization, real estate markets, and local government behavior. Land transfer and land use right sale are one of the main sources of fiscal revenue for local governments in China. Through auctions and listings, local governments can obtain revenues from land use rights, which provide funds to support local finances, infrastructure construction and public services.

To sum up, analyzing the house price from the government's point of view will play a very important role in macroeconomic regulation. Therefore, we will substitute the government's point of view and conduct regression analysis on BostonHousing dataset.

2 Dataset Background

The BostonHousing data are house prices in the Boston neighborhood in the 1970s and contain 506 observations. There are 506 observations, each containing detailed information about the house and its surroundings, and the meanings of the variables are summarized in the table below.

Name	**Meaning**
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable
nox	nitric oxides concentration
rm	average number of rooms per dwelling
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
b	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
lstat	% lower status of the population
medv	Median value of owner-occupied homes in $1000's

3 Exploratory Data Analysis

Plotting histograms of the frequency distributions for each variable revealed that medv, crim, zn, dis, and lstat showed a significant right-skewed distribution, while indus, age, rad, tax, and ptratio showed a significant left-skewed distribution.

In order to visualize the correlation between the variables, a heat map of the correlation coefficients is plotted as follows.

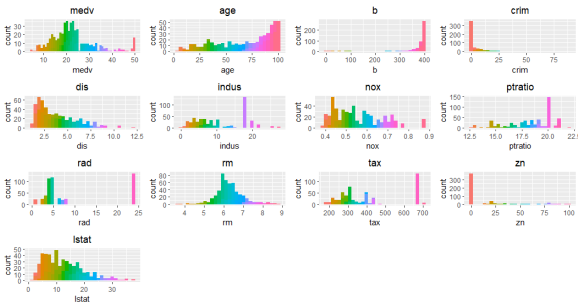

Figure 1: Heatmap of Correlation Coefficients

As can be seen from the figure, the correlation coefficients between medv and the other explanatory variables are not high. The correlation coefficients between some explanatory variables are relatively high, for example, the correlation coefficient between age and nox reaches 0.73.

4 Initial Regression Model

4.1 Data Preprocessing

We excluded the dataset outliers based on LOF values and divided the dataset into training and validation sets in 7:3.

4.2 Model Fitting

We constructed the initial model using all the variables and the results showed that the coefficients of indus, age and crim were not significant and the ANOVA analysis also concluded that they were not significant. After excluding these three variables, we ran the regression again and the R^2 of the model decreased, but the adjusted R^2 increased and all coefficients were significant.

4.3 Model Diagnostics

4.3.1 Normality, heteroscedasticity and multicollinearity

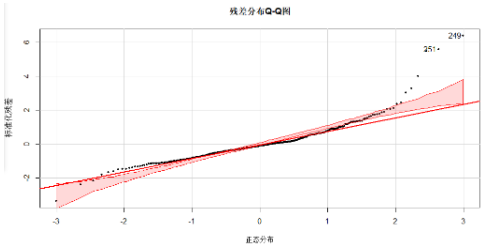

Figure 2: QQ Plot

```
        Shapiro-Wilk normality test

data:  z1
W = 0.88288, p-value = 1.721e-15
```

Figure 3: Shapiro Test

Shapiro test and QQ plot shows lack of normality.

Heteroscedasticity was detected by NCV (p-value=0.03) test but VIF shows no multicollinearity:

ZN	CHAS	NOX	RM	DIS	RAD	TAX	PTRATIO	B	LSTAT
2.17	1.05	3.65	2.12	3.32	7.87	8.60	1.84	1.40	2.68

4.3.2 Box-Cox

By optimizing the likelihood function, we obtain the Lambda optimum value of 0.26. Based on this, we perform BoxCox transformation on the data and regress the transformed data. The results show that $R^2 = 0.76$, which is a significant improvement, and the P value of Shapiro's test improves but still rejects the original hypothesis, while the P value of NCV test becomes smaller, which may be due to the data distribution being too skewed.

4.3.3 Summary

Preliminary explorations suggest that indus, age and crim are not well correlated with medv and that model residuals do not meet the assumptions of normality and homoscedasticity. In the subsequent model construction, we need to further screen the variables carefully and make the data structure conform to the model assumptions through other transformations.

4.4 Coefficients Selection

In the previous model, due to the elimination of factors such as heteroscedasticity and multicollinearity, the coefficients of *indus* and *age* variables are not significant. Therefore, it can be primarily inferred that the explanatory power of inus and age for medv is weak. Next, variable selection will be carried out through stepwise regression and full subset regression. We use the model after box-cox transformation to do the following selection.

4.4.1 Stepwise regression

Using the stepwise the stepwise regression function on the original model, the coefficients of indus and age are eliminated. The key indicators of the original and final model are compared in the table under.

	R^2	AIC
Original model	0.7988	-1352.29
Stepwise regression model	0.7986	-1355.9

4.4.2 All-subsets regression

We use the adjusted R^2 and Cp statistics as the standard for conducting full subset regression, providing appropriate references for subsequent model selection.

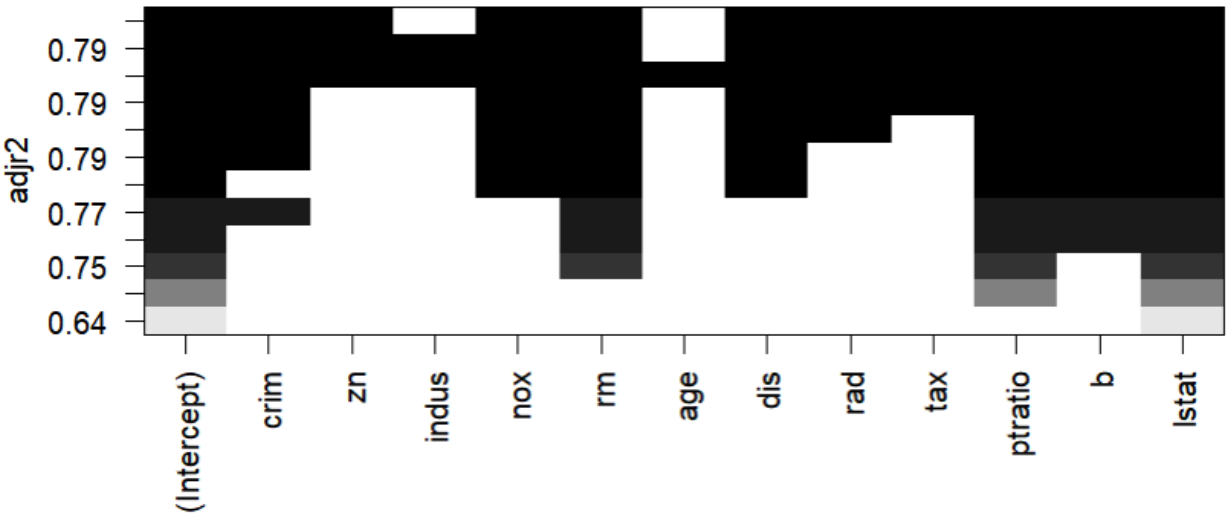

Figure 4: All Subsets Regression

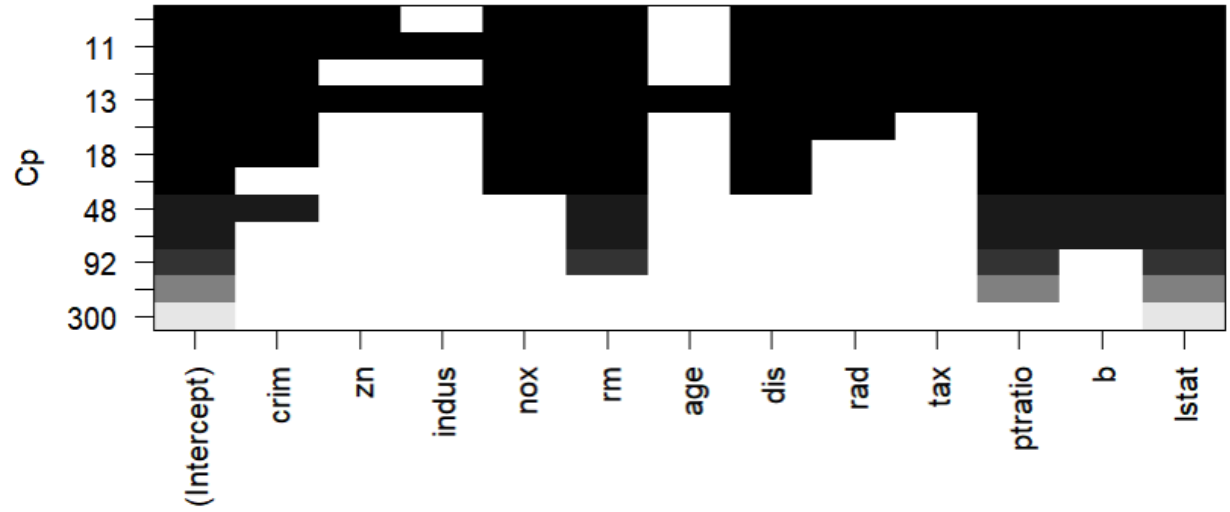

Figure 5: Cp Criterion for All Subsets Regression

Comparing the final results obtained from full subset regression and stepwise regression, it was found that both the indus and age variables were removed, which objectively indicates that these two variables cannot improve the interpretability or predictive performance of the model. Therefore, these two variables were not considered in subsequent analysis.

5 Model diagnosis and optimization

We first analyze the model given by stepwise regression. According the result summary function given, the model passed the F-test. All coefficients passed the T-test. Then we use plot function and get the following graph. From the graph, it can be seen that the residuals exhibit a more pronounced long tail distribution; the Cook distances are all within the contour line of 0.5, indicating that the model has no obvious outliers; however, the model exhibits significant heteroscedasticity.

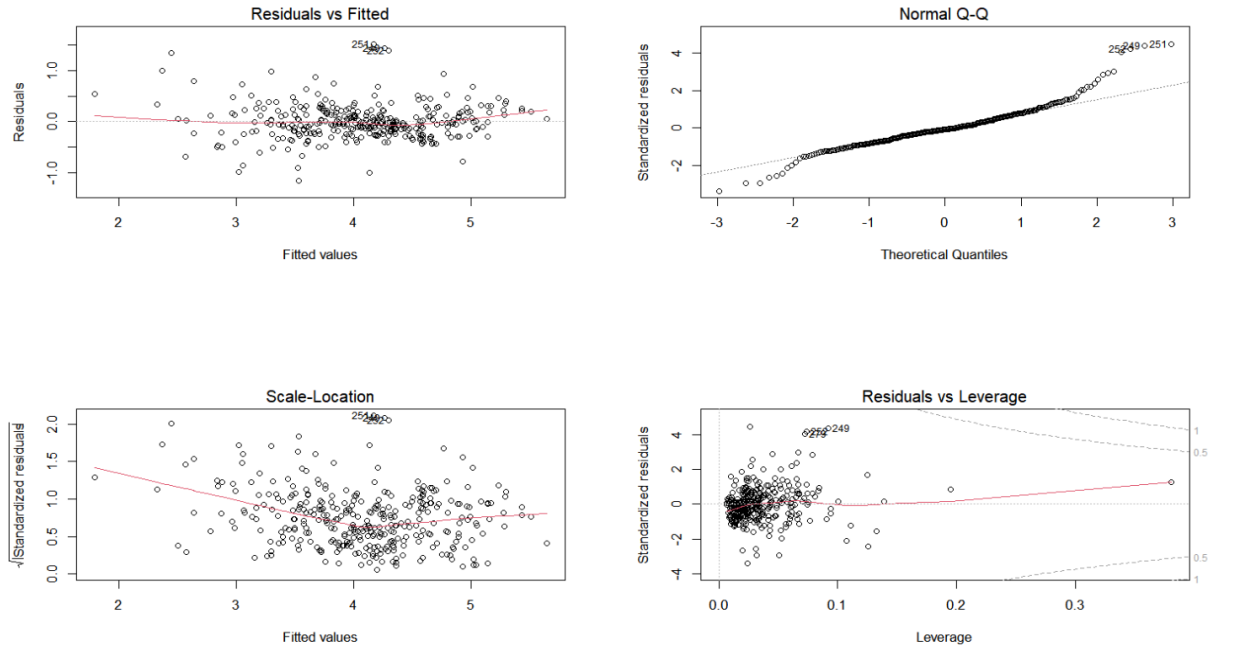

Figure 6: Diagnostics Plots for Stepwise Regression Model

Next, we use Spearman correlation test to test the correlation between the absolute value of residuals and the independent variable, and find that there is a significant correlation between the two.

5.1 Model optimization

5.1.1 WLSE

To eliminate the heteroscedasticity of the model, we will next perform WLSE. From the table above, we find the dis has the largest absolute value of spearman rho. Then we take dis To construct the weight function of WLSE. Dis shows weighted distances to five Boston employment centres. We can subjectively understand that with larger dis, the house provides a worse condition for going work, which makes the house cheaper.

After determining the independent variables for constructing the weight function, it is necessary to determine the optimal value of the power exponent m to make the regression equation optimal. We use numerical calculation method, taking m from -2.0 to 2.0 with a step size of 0.5, to calculate the value of the logarithmic maximum likelihood statistic in the regression estimation when m takes different values, and take the value corresponding to the maximum logarithmic likelihood statistic in the regression estimation.

When $m = -1.5$, the logarithmic likelihood function reaches its maximum, so the optimal value for the power exponent is -1.5. Then we get the WLSE model. The model passed the F-test. All coefficients passed the T-test. The R^2 is significantly enlarged.

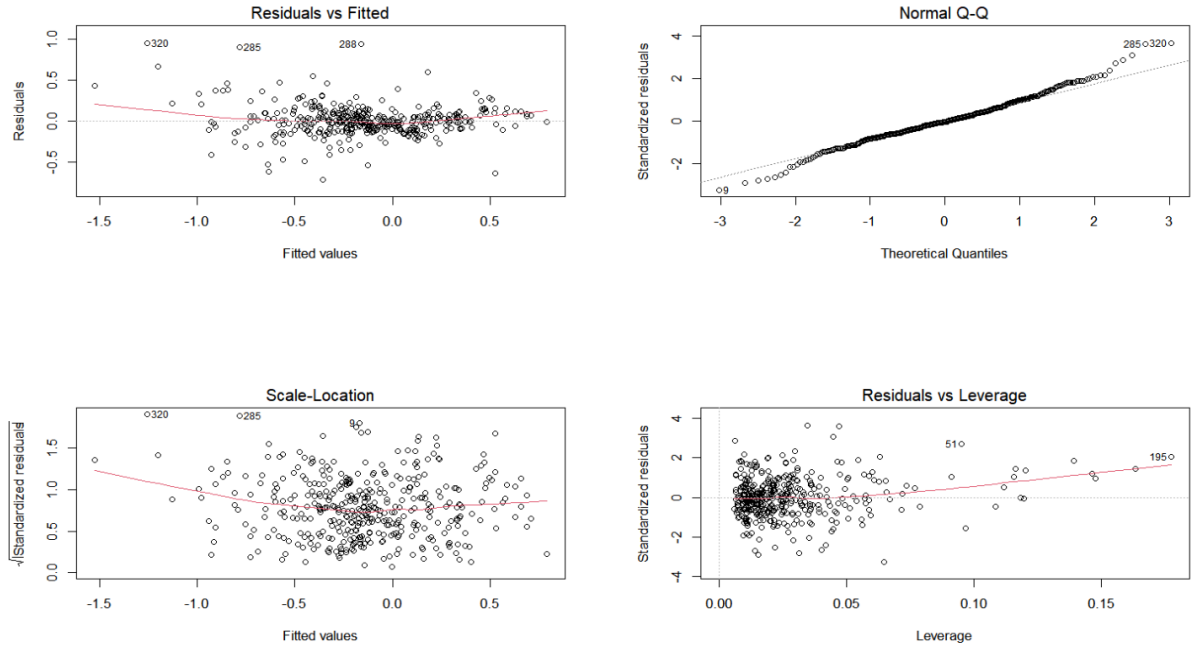

Figure 7: Diagnostic Plots for WLSE Model

5.1.2 Ridge regression

Then we deal with the multicollinearity. We first get the VIF of the original model (get from stepwise regression). The results show that there is a significant variance inflation factor for the independent variables rad and tax, and there is a certain degree of multicollinearity in the data samples. So we perform ridge regression to eliminate multicollinearity.

Firstly, we perform central standardization on the sample data of model. Then we performed ridge regression on the standardized data without intercept terms, and the results are as follows:

The ridge regression results show that the multicollinearity of the sample data in the model is weak. Using the linearRidge function in the ridge package to automatically select ridge regression parameters, we obtained a ridge regression parameter of 0.0107. The diagnostic results showed that there was no significant difference between Pr (>| t |) and original model, indicating that there was not much difference in parameter significance between this selected model obtained from ridge regression and the untreated selected model.

6 Principal Component Regression (PCR)

In consideration of the substantial number of variables and the intricacy of their interrelationships, it can be challenging to interpret them in an intuitive manner. As such, Principal Component Regression (PCR) is adopted.

6.1 Selection of Principal Components

The correlation matrix 'kmo' equals 0.83, which is greater than 0.8. This suggests strong correlations among the variables, implying that Principal Component Regression may be applied for analysis. By obtaining the eigenvalues and eigenvectors, we can then calculate the variance contribution and cumulative variance contribution rates.

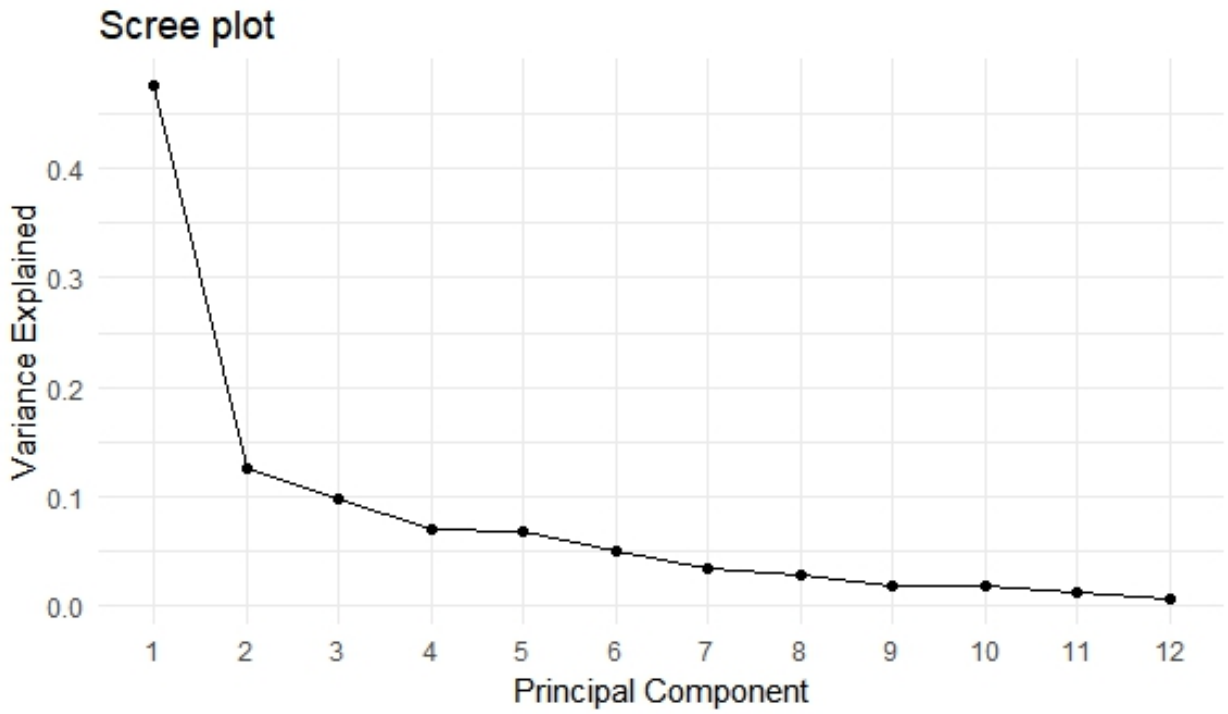

Figure 8: Scree Plot

The picture on the left displays the standard deviations, variance contribution rates, and cumulative variance contribution rates for each main component, while the scree plot is shown on the right. The left plot indicates that four principal components capture 83

As a result, we ultimately determine that there should be four principal components.

6.2 Principal Component Analysis

Having identified the four main components, we generate the factor loading matrix and observe the loading of each main component on the variables. Each type of independent variable is classified into operation component, commuting component, traffic component, and population component, based on the degree of correlation with each component to connect with reality and enhance the interpretability of the models. Possible measures that the government may adopt to adjust house prices are proposed.

Urban Environment Component: The first principal component signifies the influence of crime rate, air quality, age of houses, transport facilities, and property tax rates on house prices. These relationships suggest enhanced security, improved air quality, better housing, and a balanced tax rate could raise house prices.

Education and Diversity Component: The second component highlights the impact of education resource allocation and racial diversity on house prices. Measures for improving student-teacher ratio and resolving racial issues could enhance property values.

Socio-Economic Component: The third principal component signifies how socioeconomic factors like education quality, minority ethnicity, and lower-class population affect house prices. Elevating education and living conditions, reducing poverty could help realize higher property prices.

Location and Environment Component: The fourth component underscores how location and environmental factors influence house prices. Upgrading the urban environment, adding green spaces, enhancing education, and combating poverty could contribute to increased property values.

```
> print(loading_matrix_4)
                PC1          PC2          PC3          PC4
crim    -0.275121160 -0.32018465  0.237006524  0.118516272
zn       0.252019683 -0.32626665  0.322846046  0.246450240
indus   -0.342667195  0.12294796  0.009831742 -0.002424356
chas    -0.008021627  0.45352109  0.261174430  0.776709044
nox     -0.338790337  0.22787166  0.126962069 -0.063123792
rm       0.187081184  0.13825063  0.586525134 -0.434558845
age     -0.310832049  0.30682586 -0.013343635 -0.154018308
dis      0.318098532 -0.35033385 -0.043082983  0.226945402
rad     -0.321492115 -0.25767916  0.250784405  0.013444368
tax     -0.338586471 -0.22017476  0.195386618  0.009514167
ptratio -0.201502270 -0.30654601 -0.374561830  0.001016408
black    0.207144288  0.25808719 -0.334201037 -0.028947002
lstat   -0.306949302 -0.08285359 -0.236788198  0.229719974
```

Figure 9: Factor Loading Matrix

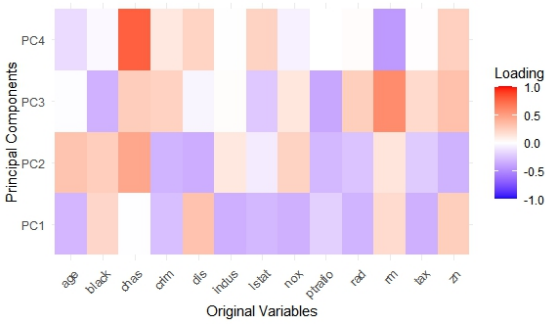

Figure 10: Biplot Displaying Variable Relationships

6.3 Model Fitting and Evaluation

Based on the extraction results of principal components in Section 5.2, we perform regression estimation again and output results. When testing the test set, the MSE of the model is found to be 0.576333403978147, which is less than the MSE of 1.179106 for the model without the application of principal component regression, indicating that principal component regression effectively optimizes the model.

7 Detection and Interpretation of Outliers and Influential Points

In practical scenarios, governments often need to accurately grasp macroeconomic trends. When studying housing prices, outliers can lead to biases in the entire model, affecting the government's subsequent analysis and decision-making. Simultaneously, the government needs to examine which observations have the most significant impact on the overall model, enabling timely strategies to be implemented in the event of large-scale market fluctuations. Therefore, it is necessary to analyze and identify influential points and outliers in the regression model, excluding certain observations, and comparing the overall effects of the model changes. Thus, an analysis of strong influential points and outliers in the regression model has been conducted.

7.1 Analysis of Strong Influential Points Regarding 'medv'

There are two main reasons for the occurrence of outliers. One is human errors, including data input errors (arising during data collection, recording, or input processes), processing errors, sampling errors, intentional outliers, etc. To identify observations that significantly affect the model's estimates, this study uses Cook's distance to determine whether a sample is a strong influential point. Initially, the criterion for identifying strong influential points was whether they exceeded $\frac{4}{n}$ (n is the sample size). However, it was found that removing points selected by this method increased the MSE, indicating that the excluded samples contained important information. Therefore, based on the threshold of $\frac{4}{n}$, the study iteratively searches for influential points that degrade the model, i.e., points whose removal reduces MSE. In the end, the study identified 277, 279, 252, 281, 268, 99, 342, 251, 295, and 272 as strong influential points. Removing these points resulted in an MSE of 0.5559, a significant improvement compared to the original model's MSE of 0.5763.

7.2 Outlier Detection Regarding 'medv'

On the basis of removing strong influential points, the study calculated the studentized residuals for each observation. Using an F-test with a significance level of 0.05, points with p-values less than 0.05 were considered outliers. However, similar to the influential point analysis, removing all outliers does not necessarily optimize the model. Based on whether the MSE after removing outliers is less than the MSE after removing strong influential points, the study identified the following outlier points: 61, 119, 263, 264, 118, 260, and 261. Removing these outlier points resulted in a final model with an MSE of 0.5534, which showed only a slight improvement compared to 0.5559, possibly because influential samples were already removed during the strong influential point analysis.

8 Summary

Analyzing the impact of housing prices on macroeconomic regulation from a government perspective will play a crucial role. On one hand, the government needs to identify the significance of regression coefficients for various variables in the model and understand their practical implications. Comparing the relative importance of different variables in influencing housing prices is essential, enabling the government to use the model for accurate predictions and assessments of housing prices in other regions.

On the other hand, the government also needs to identify instances of abnormal housing prices, preventing the outbreak of an economic bubble. Implementing measures to restrict unreasonable housing prices becomes imperative in order to maintain economic stability.

Taking into account the fitted models and the results of the principal component regression, we visualize the importance of each variable. Bar charts depicting the importance levels of variables have been created for the Weighted Least Squares Error (WLSE) model, the Box-Cox model. The y-axis represents the variable names, while the x-axis indicates the contribution to , with greater values to the right indicating higher importance.

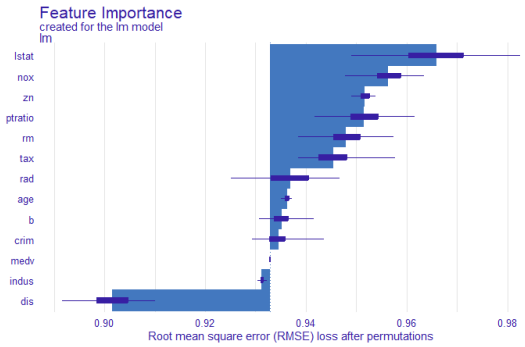

Figure 11: Bar Chart for BOXCOX

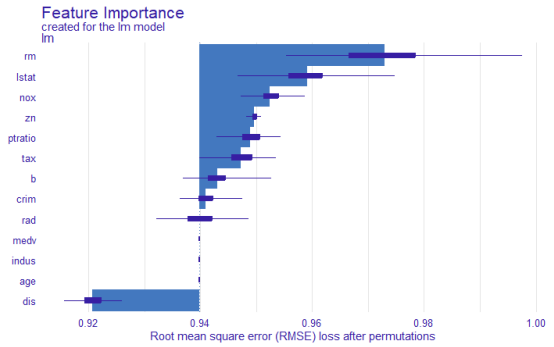

Figure 12: Bar Chart for WLSE

From the graphs, it is evident that both the 'nox' (nitric oxides concentration) and 'lstat' (lower status of the population) variables exhibit strong explanatory power in the models, with high importance when explaining 'medv' (median value of owner-occupied homes in $1000s$). The 'nox' variable, representing nitric oxide concentration, is intuitively understandable. Higher levels of nitric oxide suggest a higher likelihood of an industrial area, increased pollution, and possibly a location farther from the city center, contributing to lower housing prices.

The 'lstat' variable, indicating the percentage of lower-status population, suggests that areas with a higher concentration of lower-status residents tend to have lower housing prices. If, despite a high concentration of lower-status residents, housing prices are elevated, it may indicate a short-term surge in demand, prompting the government to increase housing supply to stabilize prices and prevent potential unrest among the lower-income population.

The 'rm' variable, representing the average number of rooms per dwelling, has a positive impact on housing prices. More rooms in a residence imply a larger area, leading to higher prices. If a region has many rooms in residential buildings but low housing prices, it may signal development issues, with an excess of unsold housing units, necessitating price reductions for sales.

The 'ptratio' variable (pupil-teacher ratio by town) is a key factor influencing housing prices and reflects the abundance of educational resources. The negative coefficients of 'ptratio' in all three models indicate that higher pupil-teacher ratios may lead to difficulties in accessing education, potentially distinguishing between properties in school districts and those outside. This explains why homes in school districts are often more expensive.

Additionally, considering transportation convenience, the 'dis' variable (weighted distances to five Boston employment centers) shows a clear negative impact on housing prices. This aligns with expectations, indicating that areas farther from employment centers may be less desirable due to longer commutes. Conversely, the 'rad' variable (index of accessibility to radial highways) suggests that areas with convenient transporta-

tion, such as proximity to radial highways, subway stations, or bus stops, are favored by consumers, resulting in higher housing prices. From a governmental perspective, there is a need for more systematic planning in the development of residential land. Creating an environment that is convenient for transportation, environmentally friendly, and comfortable and safe allows for real estate development. Additionally, providing essential infrastructure such as schools in the vicinity attracts real estate developers and encourages more consumers to make property purchases.

Furthermore, attention should be given to the adjustment of property and land tax rates. A sudden increase in property tax rates may swiftly reduce housing prices in the short term, but it could lead to panic among real estate developers, adversely affecting the long-term development of the real estate industry. An increase in land tax rates might cause consumers to hesitate in their choices, and unless housing prices are lowered to a certain extent, it may be challenging to attract a sufficient number of consumers. The government needs to strike a balance in these considerations.

In addition, there is a need to strengthen public safety measures to prevent safety concerns from deterring potential buyers and causing properties developed by real estate developers to remain unsold.

Overall, the government must carefully weigh various factors, including infrastructure development, tax rate adjustments, and public safety measures, to foster a conducive environment for real estate development and ensure the sustainable growth of the real estate industry.