Midterm Project Report

STAT3621 Statistical Data Analysis

The University of Hong Kong

Lau Yan Chun Chris
3035790941 ycclau@connect.hku.hk

1. Introduction

The Boston Housing dataset is renowned for its comprehensive range of data, encompassing various socioeconomic and demographic features that potentially impact housing prices. This study aims to leverage statistical analysis and modeling techniques to uncover insights within the dataset and shed light on the relationship between these features. Furthermore, the study focuses on developing a regression model capable of interpretating the median value of owner-occupied homes (MEDV).

2. Data Description

The dataset contains 506 sets of data with total 14 features as follows:

Feature	Type	Description
crim	continuous	Per capita crime rate by town
zn	continuous	Proportion of residential land zoned for lots larger than 25,000 square feet.
indus	continuous	Proportion of non-retail business acres per town.
chas	binary factor	Charles River dummy variable (1 if the tract bounds the river, 0 otherwise).
nox	continuous	Nitric oxide concentration (parts per 10 million).
rm	continuous	Average number of rooms per dwelling.
age	continuous	Proportion of owner-occupied units built before 1940.
dis	continuous	Weighted distances to five Boston employment centers.
rad	integer	Index of accessibility to radial highways.
tax	continuous	Full-value property tax rate per $10,000.
ptratio	continuous	Pupil-teacher ratio by town.
b	continuous	$1000(B_k - 0.63)\text{^}2$, where B_k is the proportion of Black individuals by town.
lstat	continuous	Percentage of lower status of the population.
medv	continuous	Median value of owner-occupied homes in $1000s (the target variable).

Noted that there is only 1 feature chas is binary factor. Although feature rad is integer, its value still indicates the accessibility to radial highways, it was not treated as factor in this study.

The figure 2 shows the distributions for each variable. Bar diagram was adopted for chas, all other features are presented with density diagram accordingly. Although here not every distribution of feature will be discussed in detail, it is worthy to be noted that the distribution of chas is trivially imbalance. With its imbalance characteristics, it greatly affected tasks regarding optimalization if it is selected in the scope of modeling.

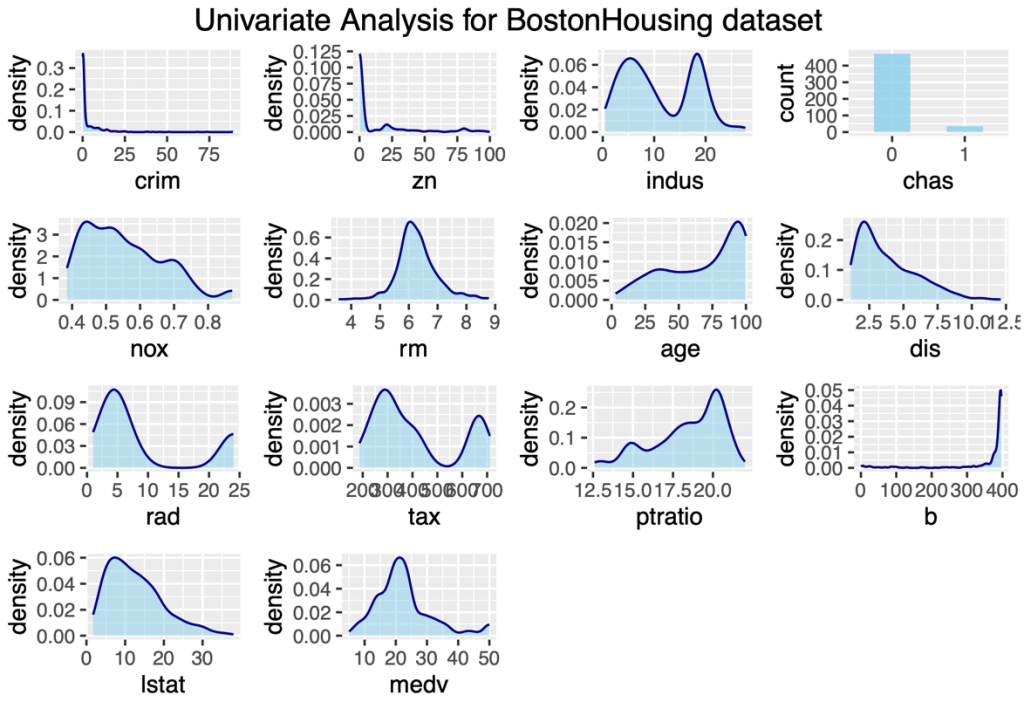

Figure 1 Density diagram for features

3. Analysis Method

3.1 Data Understanding & Interpretation

In order to gain a deeper understanding of the data and determine the significance of the features, further analysis is crucial. In this study, the relationships between features

were initially studied to obtain a brief picture of the dataset. Specific features were selected to investigate the its importance for other features. For such bivariate analysis, correlation and Wilcoxon rank sum test were mainly adopted to understand the relations.

the focus was initially placed on investigating the importance of the binary feature $chas$. To assess its impact on other features, a series of boxplots were generated to visually compare the distribution patterns based on the $chas$ factor.

3.2 Regression Model

3.2.1 Feature Selection

To construct a regression model for $medv$, a straightforward feature selection process was conducted to filter out the statistically irrelevant features. Exhaustive feature selection method was adopted given the small size of dataset.

In the process of model selection, it is crucial to consider multiple evaluation metrics to ensure a comprehensive assessment of model performance. Among different indicator, adjusted R-square R^2_{adj} was selected as a primary indicator of model goodness-of-fit. Nevertheless, Mallows' C_p, Bayesian information criterion BIC served as valuable reference points for comparative analysis.

3.2.2 Residual Analysis

It is known that different characteristics of features would be revealed in the residual diagrams, such as properties for non-linearity and non-constant variance. Simple linear regression model was applied for each feature versus $medv$ to obtain the corresponding residual diagrams, respectively. Log transformation and polynomial

transformation were introduced to attempt to solve the residual properties correspondingly.

3.2.3 Multicollinearity

Before finalizing a model, it is essential to ensure there is no undesirable multicollinearity in the model. Variance Inflation Factor (VIF) was introduced. The threshold was set as 5. In such unfortunate case, the model would be reconstructed until the multicollinearity is resolved.

3.2.4 Data Visualization

For model with multiple variables, the visualization provides insights into patterns and relationship among variables. Since the model took numbers of variables into account, Principal Component Analysis (PCA) was applied to conduct a reduction for the dimensionality of the model, enabling the model to be represented in 2D vectors.

4. Result

4.1 Data Understanding & Interpretation

4.1.1 Correlation

To understand the bivariate relations among variables, correlation table is generated and shown as figure 2. The order of features were rearranged for better presentation. For example, it reveals that there is a high correlation between rad and tax. Another interesting finding is that the $lstat$ and tax have high positive correlation, while $medv$ and tax have not.

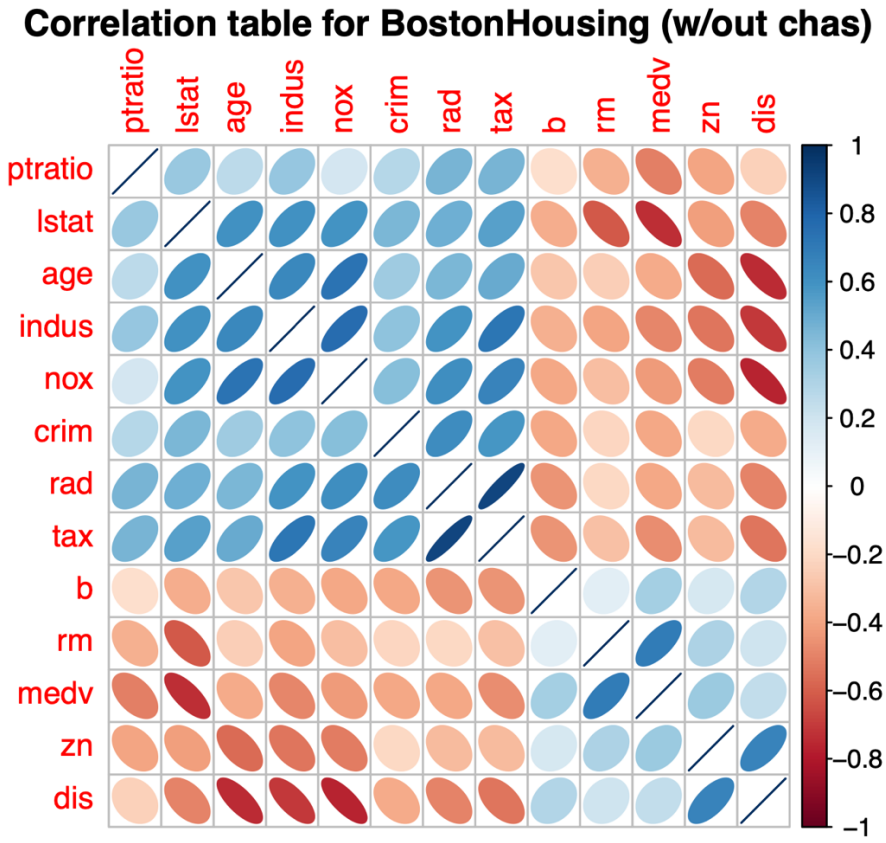

Figure 2 Correlation table for non-binary features

4.1.2 Importance of *chas*

The figure 3 demonstrate boxplots for the comparison between features by *chas*. However, by observation, there is no much meaning conclusion made.

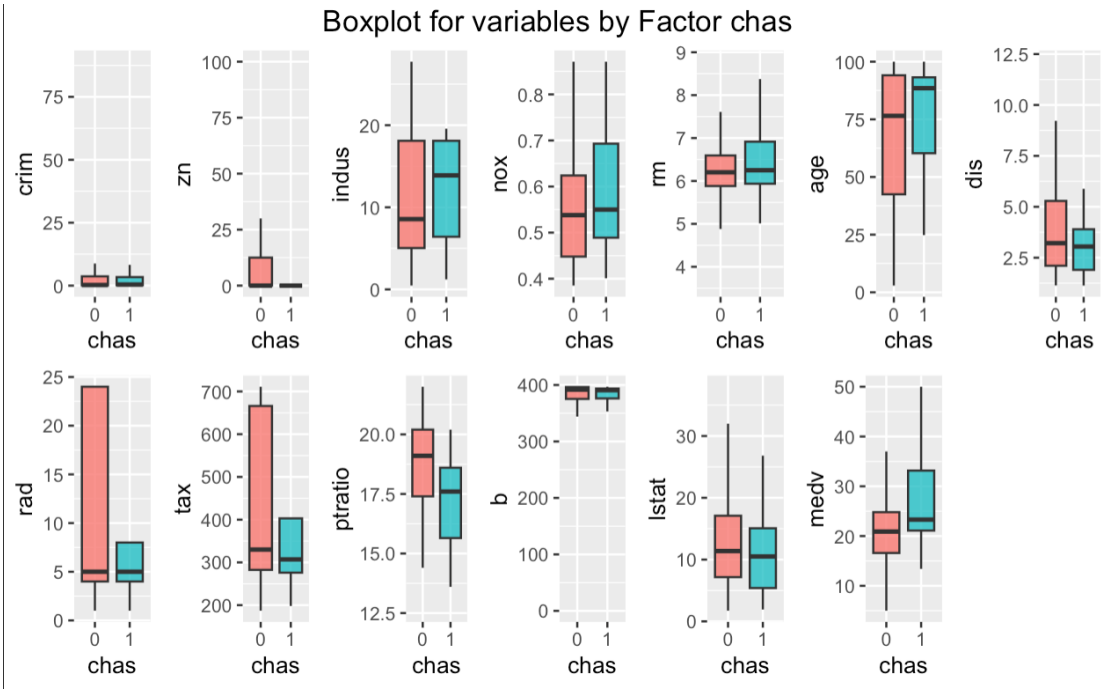

Figure 3 boxplots for features versus chas

Therefore, Wilcoxon rank sum test was introduced and there were only 3 features distributed differently by $chas$. The figure 4 shows the density comparison of $\{indus, ptratio, medv\}$ by $chas$. Few observation were made here as follows:

I. $medv \sim chas$: For people with high median value of owner-occupied homes, they tend to live next to Charles River.

II. $ptratio \sim chas$: Schools locate next to Charles River, tend to have small pupil-teacher ratio. Compared with conclusion (I).

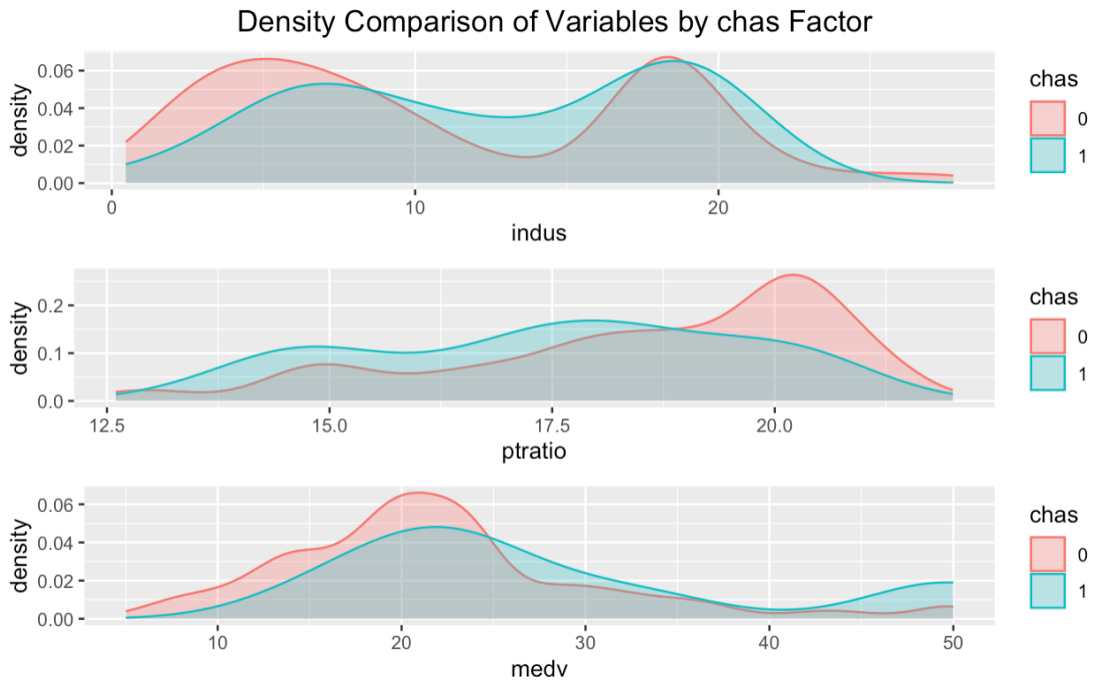

Figure 4 density diagrams for features by chas

One interesting conclusion might explain such phenomenon that people with high income may send their children to the school with small pupil-teacher ratio for better learning experience. Of course, more data and research are required to make this conclusion meaningful. Nevertheless, this finding shows the importance of $chas$ and provide a comprehensive idea of the dataset.

4.2 Regression Model

4.2.1 Residual Analysis

The residual diagrams for each feature versus $medv$ are shown in figure 5, respectively.

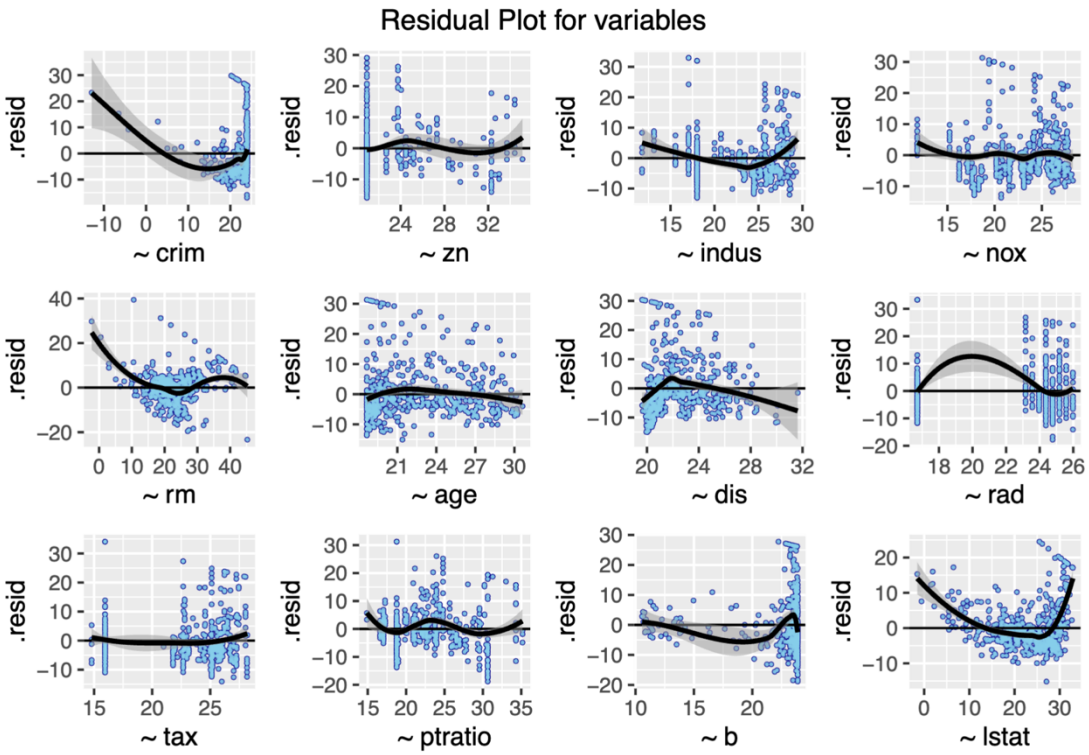

Figure 5 Residual diagram for features with SLR model

It is noted that for the residuals of $\{crim, indus, rm, lstat\}$ demonstrate a clear non-linearity property, respectively. Therefore, polynomial transformation was introduced to attempt provide better linearity for the features shown in figure {6-10}. For better visualization of the tendency, the maximum polynomial order generated was 6 as reference only. It clear demonstrates the tendency of linearity for each feature with the increment of polynomial order.

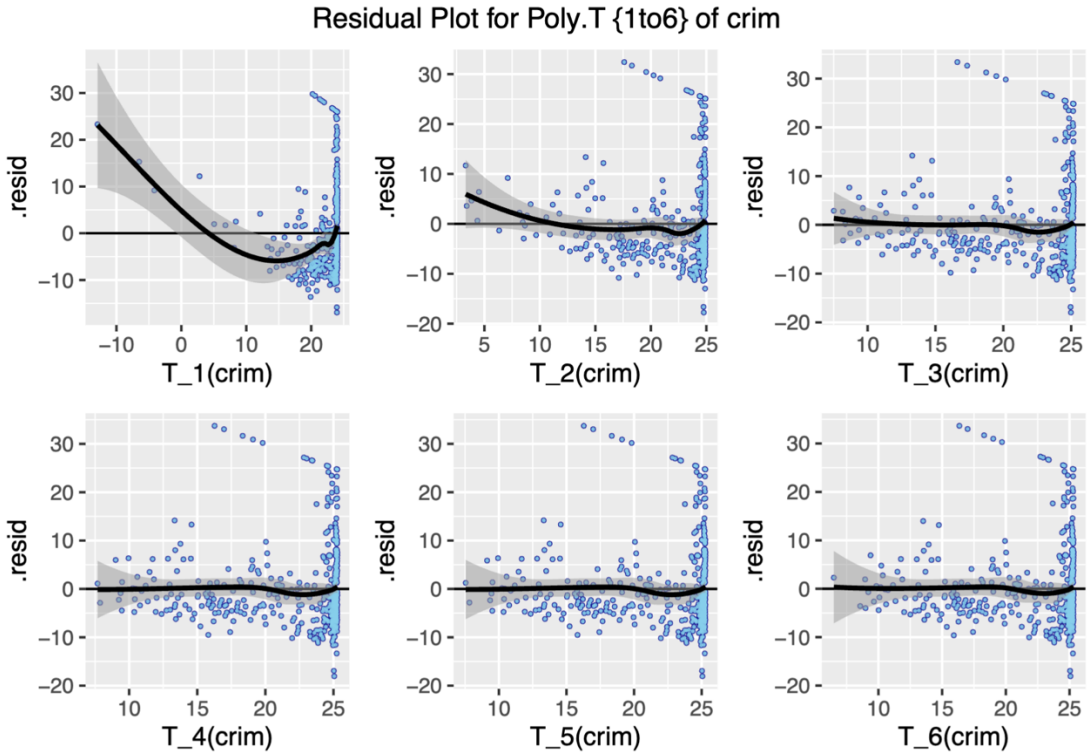

Figure 6 Residual diagram for poly. crim

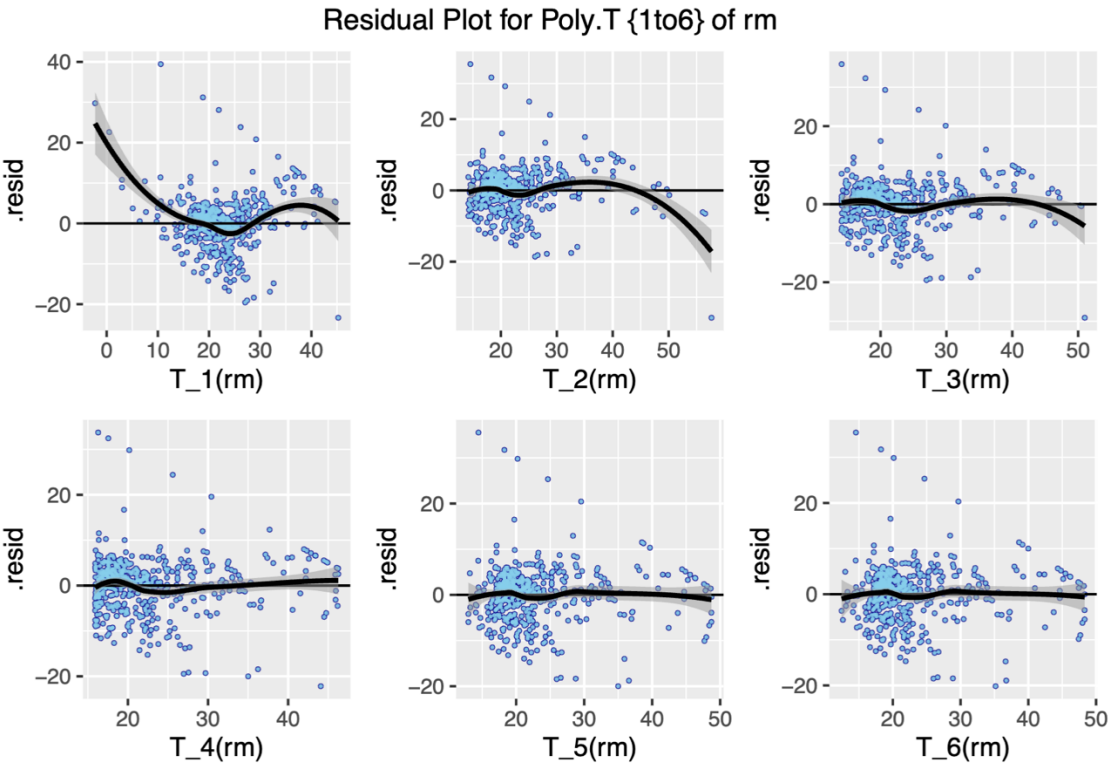

Figure 7 Residual diagram for poly. rm

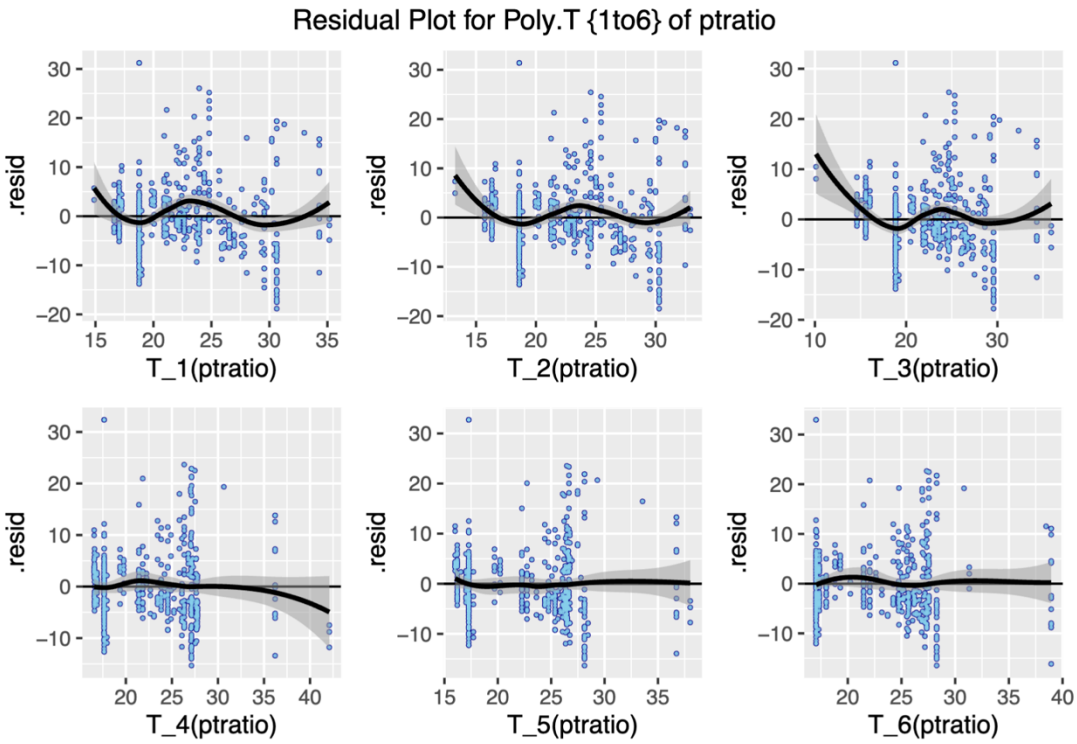

Figure 8 Residual diagram for poly. ptratio

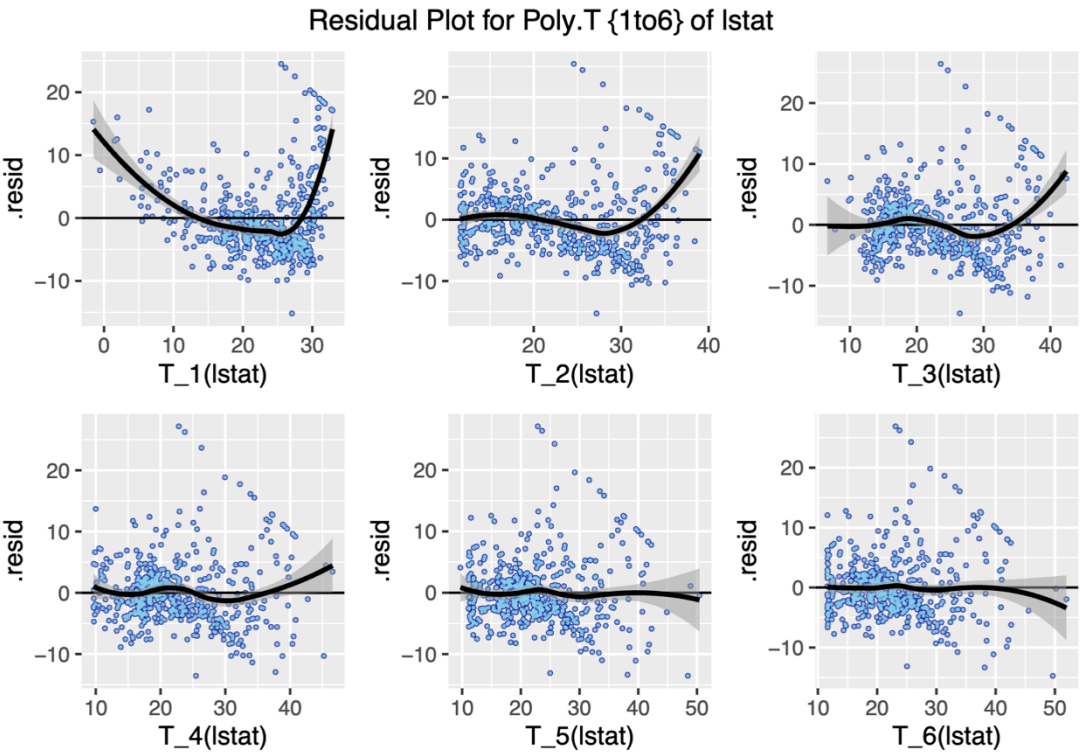

Figure 9 Residual diagram for poly. lstat

Additionally, it is observed that the residuals for $\{crim, rad, tax\}$ are dense from figure 5. Log transformation was applied to spread out the dense residuals and achieve a more uniform distribution. The residual diagrams for log transformation of $\{crim, rad, tax\}$ are shown in figure 10. The residual distributions are less dense for all three $\{crim, rad, tax\}$.

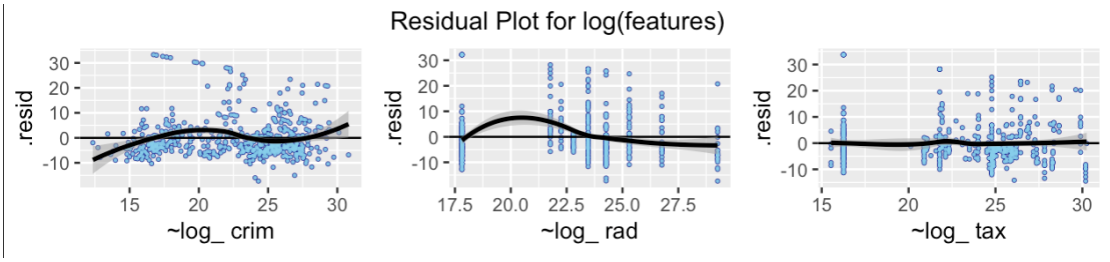

Figure 10 Residual diagram for log(features)

4.3 Feature Selection

By a straightforward forward selection with all non-transformed features, $\{indus, age\}$ were filtered out for further analysis with high p-value. After reviewing the residual analysis, the following total 22 features were considered for the regression model.

Original Feature	Feature Considered
crim	$\{\text{crim}, \text{crim}^2, \text{crim}^3\}$
zn	zn
chas	chas
nox	nox
rm	$\{\text{rm}, \text{rm}^2, \text{rm}^3\}$
dis	dis
rad	log rad
tax	log tax
ptratio	$\{\text{ptratio}, \text{ptratio}^2, \text{ptratio}^3, \text{ptratio}^4\}$
b	b
lstat	$\{\text{lstat}, \text{lstat}^2, \text{lstat}^3, \text{lstat}^4\}$

Another forward selection was processed, the model $T_{17}(X)$ suggested by R^2_{adj} with

17 features outperformed other models with value of 0.8232. The figure 11 shows the corresponding residual diagram.

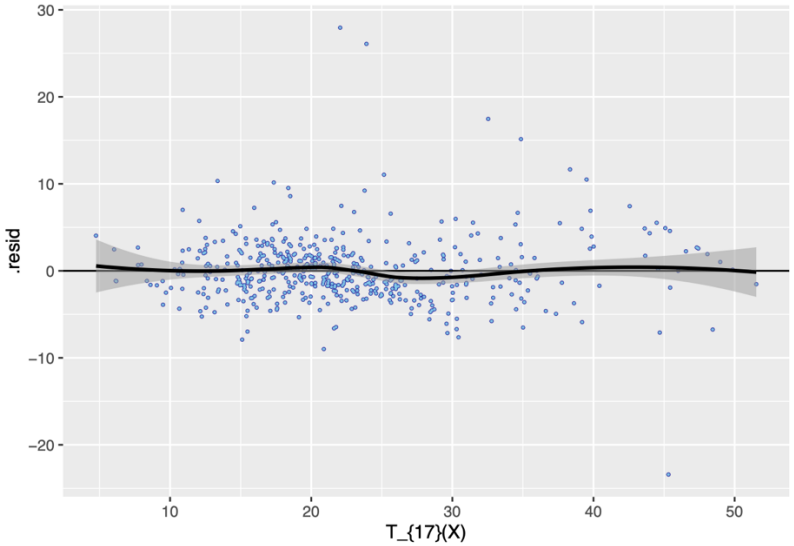

Figure 11 Residual diagram for $T_{17}(X)$ model

Compare to the residual diagram from the model $\bar{T}_{11}(X)$ only considering straightforward non-transformed features shown in figure 12, the residual analysis and feature transformations profoundly improve the model accuracy and resolve the non-linearity of the residual.

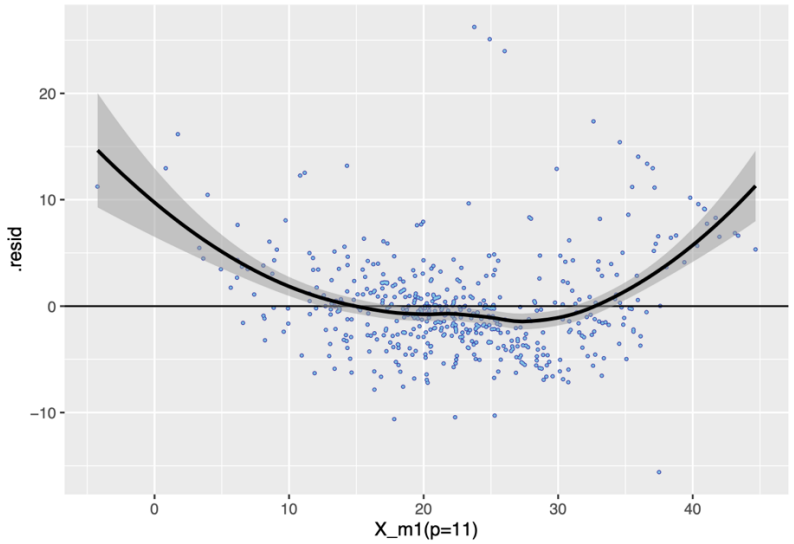

Figure 12 Residual diagram for $\bar{T}_{11}(X)$ model

4.4 Multicollinearity

The model $T_{17}(X)$ was further approved by Variance Inflation Factor with small VIF values. It indicates that there was no multicollinearity for the model.

```
print(vif)
```

```
##         crim      crim.x3         chas          nox           rm        rm.x2
## 1.564966e-01 2.582464e-01 9.144327e-01 2.387362e-01 7.893886e-03 7.702784e-03
##          dis      rad.log      tax.log   ptratio.x2   ptratio.x3   ptratio.x4
## 3.713589e-01 2.397208e-01 2.364120e-01 2.648666e-05 6.282365e-06 2.362495e-05
##            b        lstat     lstat.x2     lstat.x3     lstat.x4
## 7.224577e-01 1.643747e-03 1.519420e-04 1.016928e-04 5.567451e-04
```

4.5 Model Visualization

To visualize multivariate regression model, dimensionality reduction is required. Principal component analysis was applied to obtain the 2 normal vectors with largest variances. The figure 13 shows the result of applying PCA for the visualization of $T_{17}(X)$.

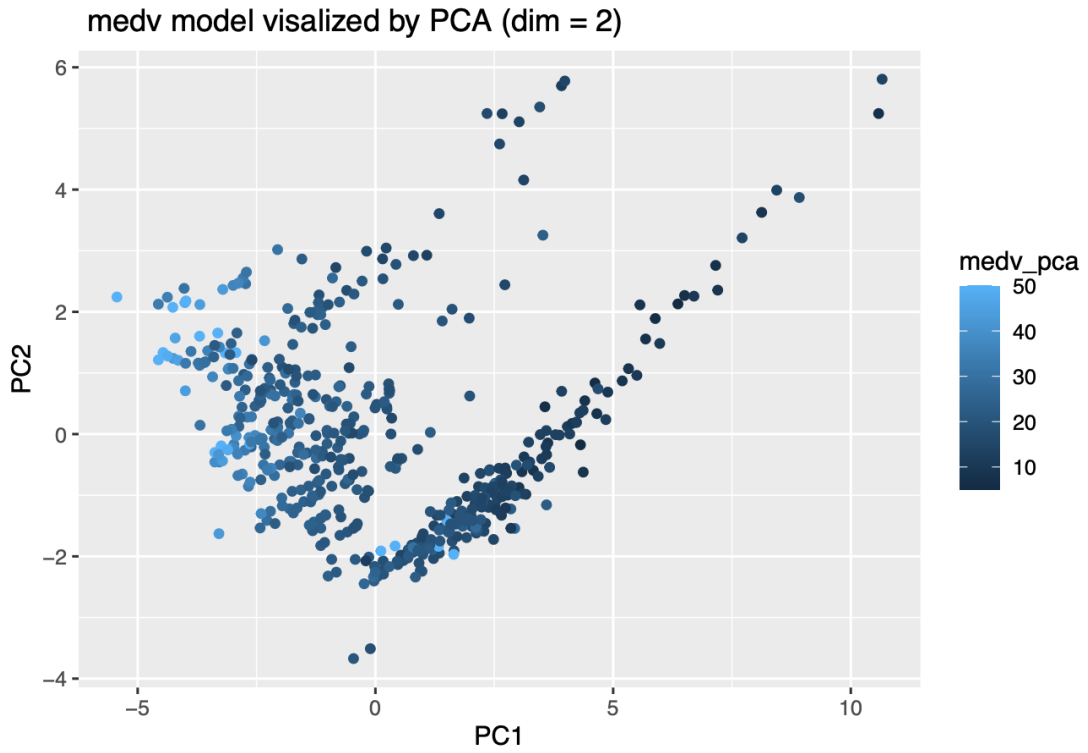

Figure 13 PCA plot for $T_{17}(X)$ model

5. Discussion

Although the result of constructing regression model for $medv$ is satisfactory, this model is not suitable to predict the value of $medv$ given any new set of features. From the point of view of machine learning, there is no validation policy to regulate the model. No one would have any indicator to know whether it is overfitted or not. Nevertheless, this model, still, is able to play a key role to understand the relationship between $medv$ and other features in Boston Housing dataset.

6. Conclusion

In conclusion, this study presented a comprehensive analysis of the Boston Housing dataset, focusing on the development of a regression model to understand and interoperate the median value of owner-occupied homes $medv$, Through data understanding and interpretation, valuable insights were found. Residual analysis was conducted to assess the model's assumptions and address non-linearity and non-constant variance using log transformation and polynomial transformation. Multicollinearity was examined using the Variance Inflation Factor (VIF), ensuring the model's stability. Finally, data visualization using Principal Component Analysis (PCA) provided visual representations of the model in reduced-dimensional space.