Multiple Regression Analysis of Boston Housing Values

Elena Bateman, Cherry Li, Jonathan McGee, Jeanna Kim, Gloria Thet

1. Introduction

Boston is the capital and largest city in Massachusetts, located in northeastern United States. More than one-fourth of the city is water, including the Charles River, Boston Harbor, and part of the Atlantic Ocean. In the colonial times, Boston was a popular stop on a major trade route. During the late 18th century, Boston turned to transform to an urban landscape as the population grew and they needed more space. In our study, we are interested in answering the following research question: Can median housing prices in the Boston area be predicted from different geographic and demographic factors?

The data we used reports on different housing values in the suburbs of Boston. It is extracted from the "Boston" data set found in the "MASS" R package https://www.r-project.org. The data contains information about per capita crime rate by town, proportion of residential land zone for lots over 25,000 square feet, proportion of non-retail business acres per town, if a tract bounds the Charles River, the nitrogen oxides concentration, the average number of rooms per dwelling, proportion of owner-occupied units, weighted mean of distances to five Boston employment centers, the index of accessibility to highways, property tax rates, pupil-to-teacher ratio, proportion of African Americans in town, lower status of the population, and the median value of owner-occupied homes.

Since the median value of Boston housing is dependent on a variety of factors, a simple linear regression model would not be able to fully capture and explain the variation in our response variable with only one explanatory variable. An expansion of simple linear regression is multiple linear regression, which allows us to analyze the relationship between several different explanatory variables and our response variable. In this study, we implemented a multiple linear regression model to predict the median value of owner-occupied homes with the knowledge of twelve predictor variables. The population linear model is as follows:

$$Y = \beta_0 + \beta_1(crim) + \beta_2(indus) + \beta_3(chas) + \beta_4(nox) + \beta_5(rm) + \beta_6(age) + \beta_7(dis) + \beta_8(rad) + \beta_9(tax) + \beta_{10}(ptratio) + \beta_{11}(black) + \beta_{12}(lstat) + \varepsilon$$

In the following report, section 2 summarizes the statistics and distributions for each individual variable in our model as well as the relationship between variables. Section 3 offers three potential predictive models, from which we have chosen one as the most appropriate model in predicting the median value of Boston housing. This section also consists of an analysis of the validity of our model as well as an interpretation of our results. Lastly, we discuss the consistencies between our results and previous research on this topic along with potential limitations and areas for future improvement in our analysis.

2. Data Description

Our data comes from the "Boston" dataset in the R-Studio package "MASS." The Boston Housing Dataset is derived from information from the US Census Service concerning housing factors in the area of Boston, Massachusetts. The dataset includes columns:

CRIM: the per capita crime rate by town

ZN: the proportion of residential land zoned for lots over 25,000 square feet

INDUS: the proportion of non-retail business acres per town

CHAS: the Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX: the nitric oxide concentration rate in parts

PTRATIO: pupil-teacher ratio by town

LSTAT: % lower status of the population

RM: the average number of rooms per dwelling

AGE: the proportion of owner-occupied units built prior to 1940

DIS: the weighted distances to five Boston employment centres

RAD: the index of accessibility to radial highways

TAX: full-value property-tax rate per $10,000

BLACK: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

MEDV: Median value of owner-occupied homes in $1000's

Within the dataset, there are a total of 506 observations recorded and 14 variables. For our model, we chose to use median home value (MEDV) as the response variables, using predictors crime (CRIM), industrial acres proportion (INDUS), location in reference to the Charles River (CHAS), nitric oxide concentration rate (NOX), room number (RM), proportion of home age (AGE), distance to Boston employment centers (DIS), accessibility to radial highways (RAD), tax rate (TAX), pupil-teacher ratio (PTRATIO), proportion of black individuals (BLACK), and economic status (LSTAT); excluding the variable ZN from our model for simplicity in transformation. Of the variables, only room (RM) appears to be normally distributed, as reflected in Figure A.2 and the variables' respective summary statistics in Table A.3. The variables also don't appear to be overly correlated, as shown in Table A.4. However, analyzing the scatterplot matrix, many of the relationships between variables do not appear linear.

3. Results and Interpretation

Our initial model is the full model with all twelve original predictor variables. Although the preliminary model has a relatively high R^2 value of 73.46% and the overall model is significant according to its low p-value for the F-test, these summary statistics cannot be considered reliable unless the model assumptions are met. Through a thorough analysis of the full model's diagnostic plots, we have concluded that the model is not valid. As can be seen in Figure A.5, the upward concave line in the Residuals against Fitted Values plot provides evidence against the linearity of the relationship assumption. In fact, a closer look at the scatter plot matrix shows that many predictor variables lack an evident linear relationship with the response variable. Additionally, the Normal Q-Q plot suggests that the errors are consistent with a heavy-tailed distribution, failing to satisfy the normality of the error term assumption. The U-shaped pattern in the Residuals vs Fitted and Scale-Location plots also implies that the error term does not have constant variance, resulting in a moderate violation of homoscedasticity. Lastly, we observed that there are several outliers, leverage points, and influential points that may require further inspection. Based on these observations, we then consider a transformation of our numerical variables.

Using the Box-Cox method, we apply appropriate power transformations to each of our numerical variables based on the recommended "Rounded Pwr." For the sake of simpler and more meaningful interpretations, we have further rounded the recommended λ for multiple variables. Thus, our second model is the transformed model with all variables. Although this model has shown a significant improvement in validating model assumptions (Figure A.6), the large number of predictor variables raises our concerns regarding overfitting, which reduces the predictive ability of our model. Through an inspection of the Variable Inflation Factors (VIF) for each predictor, we first observe if there is a significant issue of multicollinearity. In fact, nearly half of our variables have a VIF value of greater than

5, which suggest a relatively high amount of correlation between the variables. Furthermore, we have examined the effect of each predictor on the response variable through the Added Variable Plots in Figure A.7. Many predictors seem to add little to the prediction of the median value of owner-occupied homes. With these factors in mind, we consider performing variable selection to find the best subset model.

Using the forward and backward stepwise regression methods, we find that the removal of the crime and industry variables will produce a more appropriate model with a greater adjusted R^2 and lower AIC. In comparing the goodness of fit criteria between the three models, Table 3.1 shows that the third (final) model has the highest adjusted R^2 as well as the lowest AIC and BIC. This justifies our decision to choose the final model as the most appropriate model in predicting the median value of owner-occupied homes in Boston.

	AIC <dbl>	AdjR2 <dbl>	BIC <dbl>
Final model	625.0590	0.7756	671.5509
Transformed model	627.3656	0.7754	682.3106
Initial model	3037.2345	0.7282	3096.4060

Table 3.1 Model Comparison

In performing a partial F-test with the final (reduced) model as the null hypothesis and the transformed (full) model as the alternative hypothesis, we find a p-value of 0.4375, which is greater than 0.05 (refer to Figure A.8). Thus, we fail to reject the null hypothesis, meaning that there is not sufficient evidence against the final model in favor of the reduced model. Furthermore, to analyze the validity of the final model, we can analyze the diagnostic plots.

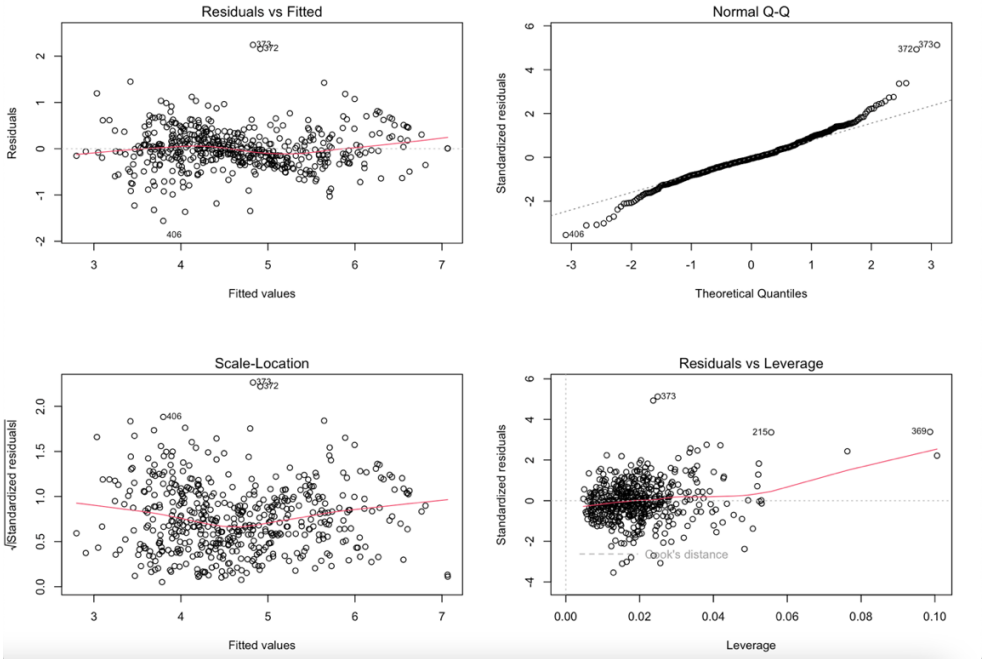

Fig. 3.2 Diagnostic Plots of Final Model

As can be seen, the Residuals vs Fitted and Scale-Location plots have greatly improved, so the linearity of the relationship assumption along with the constant variance of the error term assumption now hold. According to the Normal Q-Q plot, the points are now more aligned to the straight line, except near the ends, so we can consider the normality of the error term assumption to be satisfied. Although there are

still several outliers, leverage points, and influential points, the observations' Cook's distances seem to have decreased. Based on these observations, our third model is valid.

Thus, our final model has the following regression equation:

$$\widehat{\sqrt{medv}} = 6.116 + 0.465\left(\frac{1}{nox}\right) + 0.228(rm) + 2.156 \times 10^{-3}(age) - 0.461(\log(dis)) + 0.116\left(\sqrt{rad}\right) - 0.058\left(\sqrt{tax}\right) - 3.069 \times 10^{-6}(ptratio^4) + 8.725 \times 10^{-12}(black^4) - 0.990(lstat)$$

Our final model has a R^2 value of 77.96%, meaning that nearly 78% of the variation in the median value of owner-occupied homes can be explained by the regression model. Additionally, the F-test for overall model significance has a p-value of below 2.2×10^{-16}, which is less than 0.05, so we have enough evidence to reject the null hypothesis and conclude that our model is significant. Our model predicts the following:

1. For a 1 unit increase in the inverse of the nitrogen oxides concentration, we expect on average a 0.465 unit increase in the square root of the median value of owner-occupied homes.
2. For a 1 unit increase in the average number of rooms per dwelling, we expect on average a 0.228 unit increase in the square root of the median value of Boston owner-occupied homes.
3. For a 1 unit increase in the proportion of owner-occupied units built prior to 1940, we expect on average a 2.156×10^{-3} unit increase in the square root of the median value of owner-occupied homes.
4. For a 1 unit increase in the log of the weighted mean of distances to five Boston employment centers, we expect on average a 0.461 unit decrease in the square root of the median value of owner-occupied homes.
5. For a 1 unit increase in the square root of the index of accessibility to radial highways, we expect on average a 0.465 unit increase in the square root of the median value of owner-occupied homes.
6. For a 1 unit increase in the square root of the full-value property-tax rate per $10,000, we expect on average a 0.058 unit decrease in the square root of the median value of owner-occupied homes.
7. For a 1 unit increase in the pupil-teacher ratio to the fourth power, we expect on average a 3.069×10^{-6} unit decrease in the square root of the median value of owner-occupied homes.
8. For a 1 unit increase in the value $(1000(Bk - 0.63)^2)^4$, where Bk is the proportion of black individuals, we expect on average a 8.725×10^{-12} unit increase in the square root of the median value of owner-occupied homes.
9. For a 1 unit increase in the percent of the lower status of the population, we expect on average a 0.990 unit decrease in the square root of the median value of owner-occupied homes.

4. Discussion

In this study, the data on various geographic and demographic features of Boston is studied through a multiple linear regression model to predict the median value of owner-occupied homes in Boston. Based on our analysis, an appropriate predictive model includes the following predictor variables: nitrogen oxides concentration, average number of rooms per dwelling, proportion of owner-occupied units built prior to 1940, weighted mean of distances to five Boston employment centers, accessibility to radial highways, full-value property-tax rate per $10,000, pupil-teacher ratio by town, proportion of African Americans by town, and percent of lower status of the population. From our results, the percent of the lower status of the population along with the average number of rooms per dwelling appear to be the variables that most influence the median value of owner-occupied homes in Boston.

In the real world, our model would make sense. For the variable nox, it makes sense that it has a decreasing impact on the median housing values because a high concentration of nitrogen oxide is not healthy to live in. It also makes sense that more rooms in a house would increase the median housing values. It is also understandable that there is a small increase in median housing values for an increase in the proportion of owner-occupied units and that the further away from Boston employment centers the house is, the smaller the median housing value. It would make sense that the closer the house is to highways, the median housing value would also decrease and for an increase in property tax rate, the median value of Boston housing would decrease. A higher pupil-teacher ratio would decrease housing prices. Finally, an increase in the percentage of lower status of population would decrease the median value of housing in Boston. Recently, a similar study has been completed predicting home values using machine learning algorithms. The study uses data transformation, model selection and comparison, and modeling of similar variables to predict home values for sites such as Zillow. Although our methods differ, their report shows many consistencies between their results and our results.

The main limitation of our analysis is the interpretation of the coefficients since there is a variety of transformations among the response variable and the predictors (e.g. log, square root, inverse, to the 4th power). Additionally, since our variables consist of a variety of units, it is more difficult to directly compare the level of influence that each predictor variable has on our response variable through their estimated coefficients. Another limitation of our analysis is that not all potential influential factors are taken into account. The selection of predictor variables is based on the variables provided in the Boston dataset within the MASS package. In the future, we can consider adding additional variables that we believe would be significant in predicting the median value of homes. This data can be extracted and cleaned from the Massachusetts Online Database provided by Vision Government Solutions. Lastly, we have not accounted for potential variation within each town so it would be difficult to make accurate predictions. In a future research study, we can consider analyzing and creating a multiple linear regression model to predict the appraised value of houses in a single town of Boston.

References

1. Harrison, D. and Rubinfeld, D.L. (1978) “Hedonic prices and the demand for clean air.” *J. Environ. Economics and Management* 5, 81–102.
2. Belsley D.A., Kuh, E. and Welsch, R.E. (1980) “Regression Diagnostics.” *Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
3. Jermain, Nate. (2019) “Home Value Prediction.” *Towards Data Science*. https://towardsdatascience.com/home-value-prediction-2de1c293853c.

Appendix

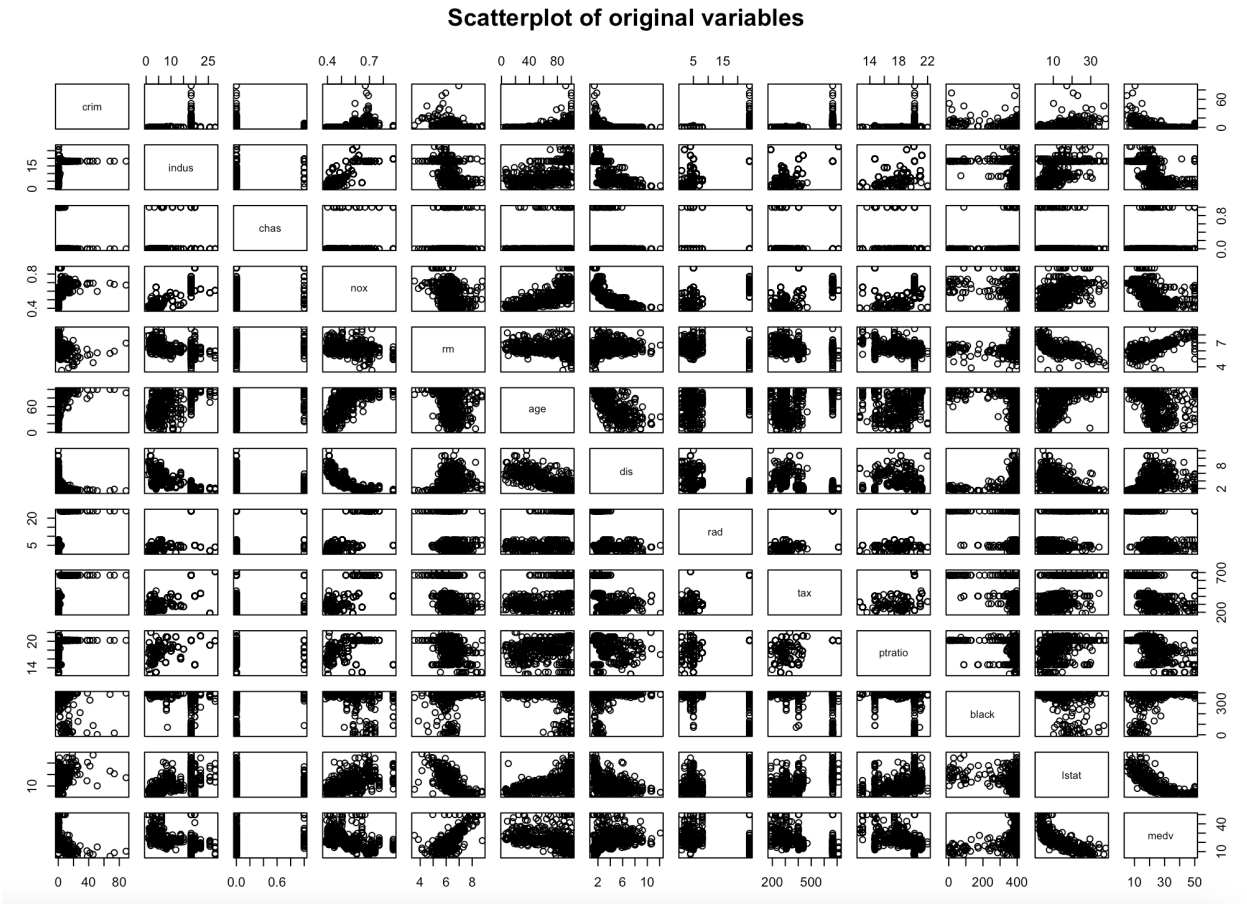

Fig. A.1 Scatter Plot Matrix of Original Variables

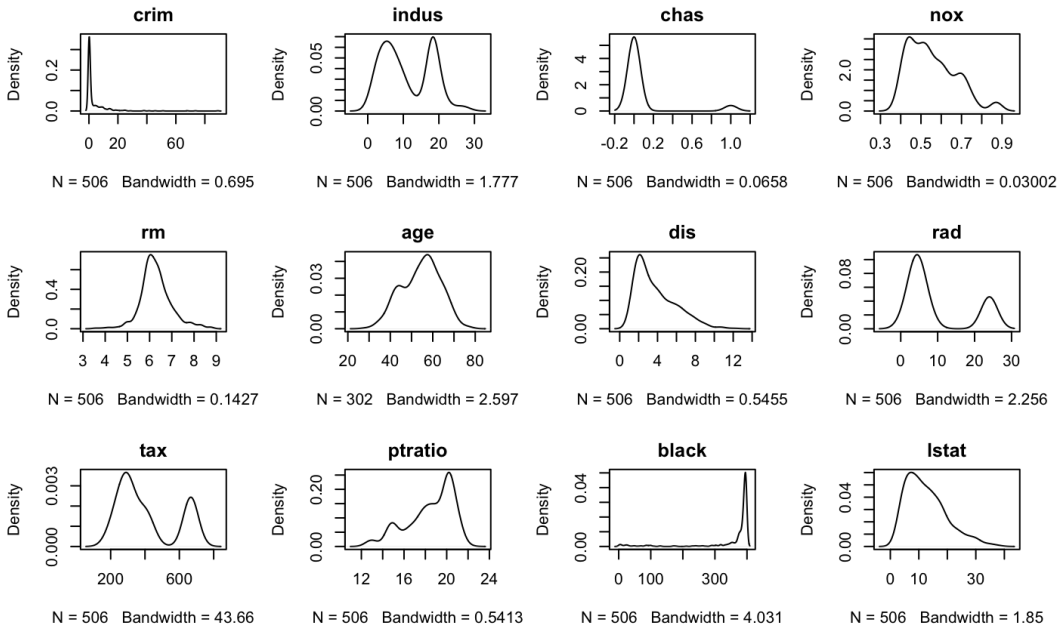

Fig. A.2 Distributions of Original Variables

	Mean <dbl>	Median <dbl>	Variance <dbl>
crim	3.61352356	0.25651	7.398658e+01
indus	11.13677866	9.69000	4.706444e+01
chas	0.06916996	0.00000	6.451297e-02
nox	0.55469506	0.53800	1.342764e-02
rm	6.28463439	6.20850	4.936709e-01
age	68.57490119	77.50000	7.923584e+02
dis	3.79504269	3.20745	4.434015e+00
rad	9.54940711	5.00000	7.581637e+01
tax	408.23715415	330.00000	2.840476e+04
ptratio	18.45553360	19.05000	4.686989e+00

Table A.3 Variable Summary Statistics

	crim	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
crim	1.0000	0.4066	-0.0559	0.4210	-0.2192	0.3527	-0.3797	0.6255	0.5828	0.2899	-0.3851	0.4556	-0.3883
indus	0.4066	1.0000	0.0629	0.7637	-0.3917	0.6448	-0.7080	0.5951	0.7208	0.3832	-0.3570	0.6038	-0.4837
chas	-0.0559	0.0629	1.0000	0.0912	0.0913	0.0865	-0.0992	-0.0074	-0.0356	-0.1215	0.0488	-0.0539	0.1753
nox	0.4210	0.7637	0.0912	1.0000	-0.3022	0.7315	-0.7692	0.6114	0.6680	0.1889	-0.3801	0.5909	-0.4273
rm	-0.2192	-0.3917	0.0913	-0.3022	1.0000	-0.2403	0.2052	-0.2098	-0.2920	-0.3555	0.1281	-0.6138	0.6954
age	0.3527	0.6448	0.0865	0.7315	-0.2403	1.0000	-0.7479	0.4560	0.5065	0.2615	-0.2735	0.6023	-0.3770
dis	-0.3797	-0.7080	-0.0992	-0.7692	0.2052	-0.7479	1.0000	-0.4946	-0.5344	-0.2325	0.2915	-0.4970	0.2499
rad	0.6255	0.5951	-0.0074	0.6114	-0.2098	0.4560	-0.4946	1.0000	0.9102	0.4647	-0.4444	0.4887	-0.3816
tax	0.5828	0.7208	-0.0356	0.6680	-0.2920	0.5065	-0.5344	0.9102	1.0000	0.4609	-0.4418	0.5440	-0.4685
ptratio	0.2899	0.3832	-0.1215	0.1889	-0.3555	0.2615	-0.2325	0.4647	0.4609	1.0000	-0.1774	0.3740	-0.5078
black	-0.3851	-0.3570	0.0488	-0.3801	0.1281	-0.2735	0.2915	-0.4444	-0.4418	-0.1774	1.0000	-0.3661	0.3335
lstat	0.4556	0.6038	-0.0539	0.5909	-0.6138	0.6023	-0.4970	0.4887	0.5440	0.3740	-0.3661	1.0000	-0.7377
medv	-0.3883	-0.4837	0.1753	-0.4273	0.6954	-0.3770	0.2499	-0.3816	-0.4685	-0.5078	0.3335	-0.7377	1.0000

Table A.4 Variable Correlation

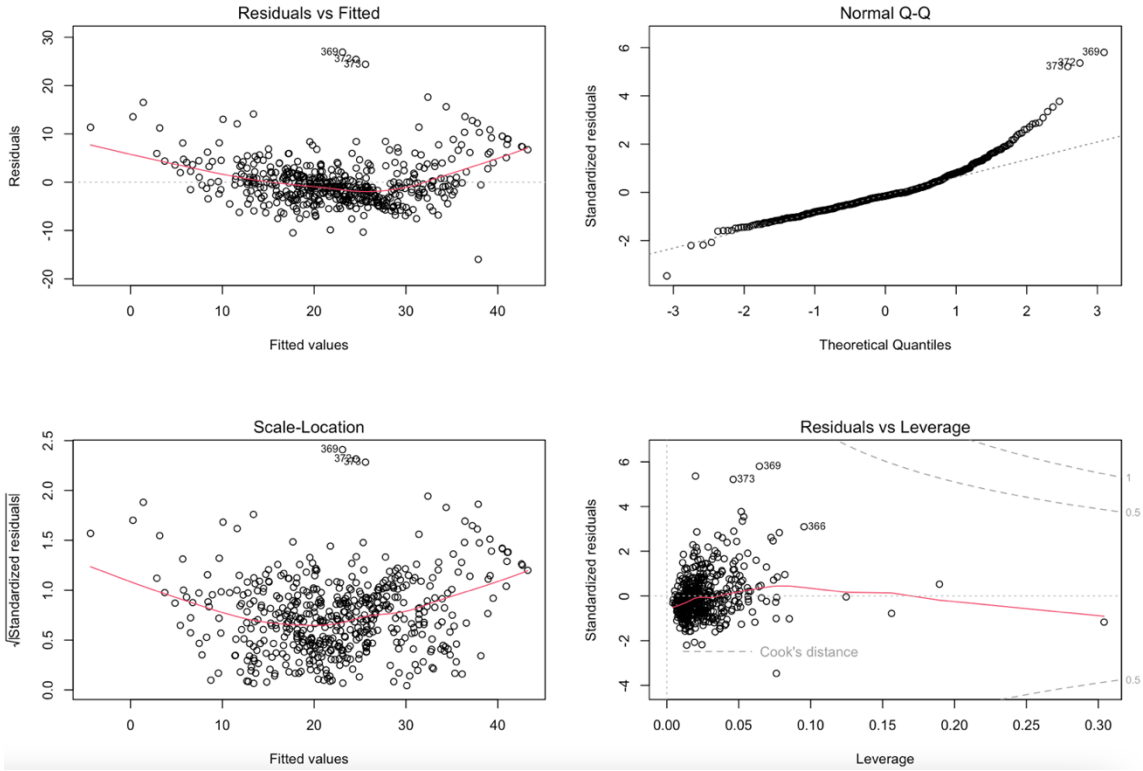

Fig. A.5 Diagnostic Plots of Full Model

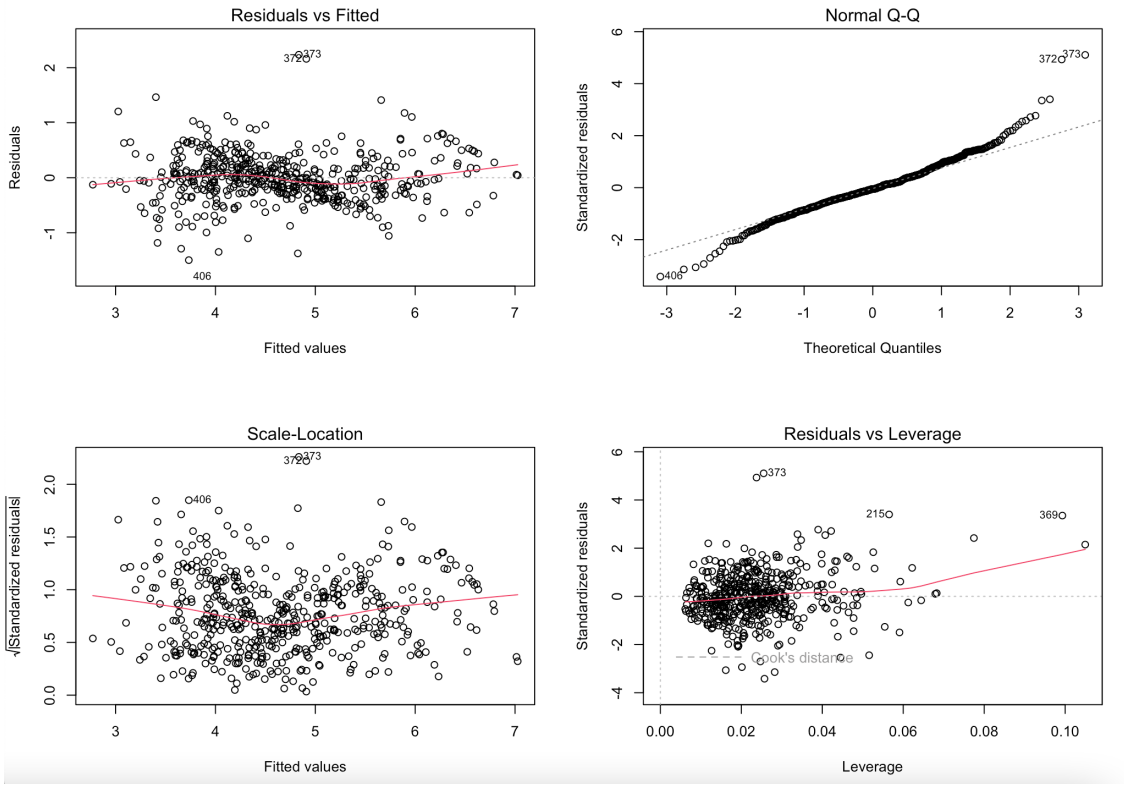

Fig. A.6 Diagnostic Plots of Transformed Model

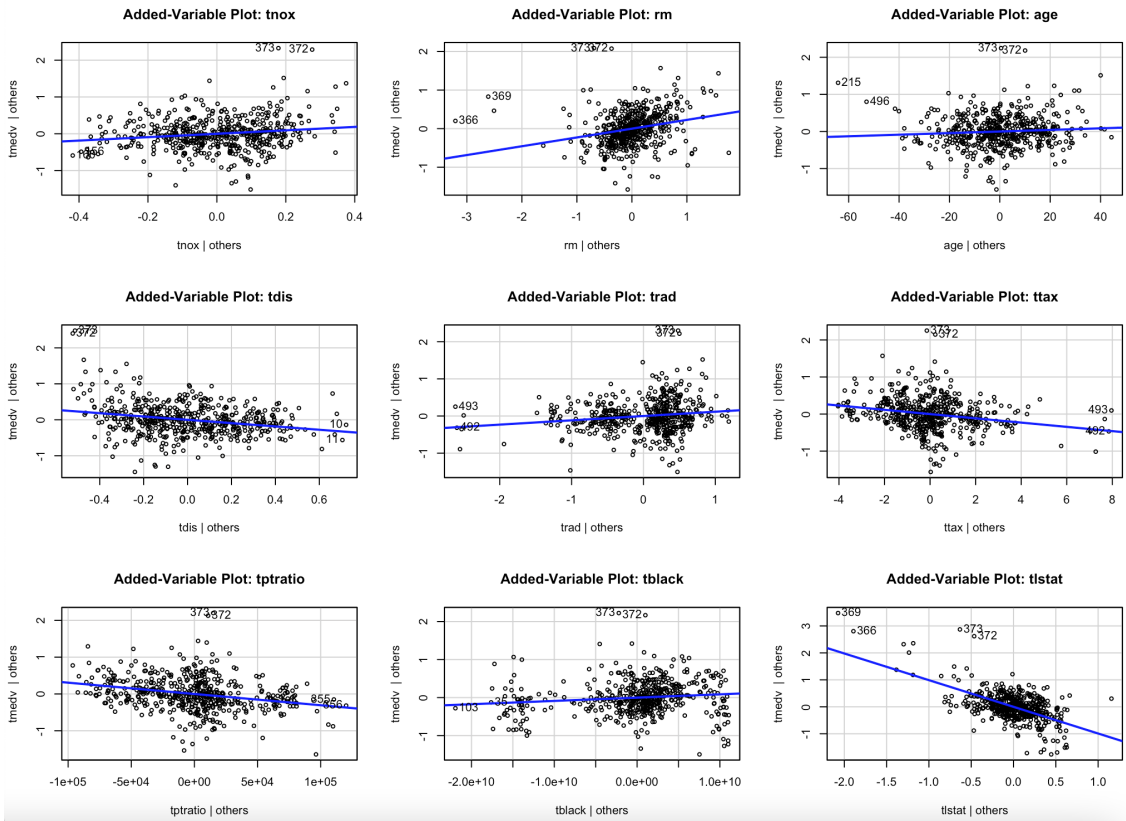

Fig. A.7 Added-Variable Plots of Final Variables After Transformation

```
Analysis of Variance Table

Model 1: tmedv ~ tnox + rm + age + tdis + trad + ttax + tptratio + tblack +
    tlstat
Model 2: tmedv ~ tcrim + tindus + tnox + rm + age + tdis + trad + ttax +
    tptratio + tblack + tlstat
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    496 97.561
2    494 97.235  2   0.32595 0.828 0.4375
```

Fig. A.8 R Output for Partial F-test

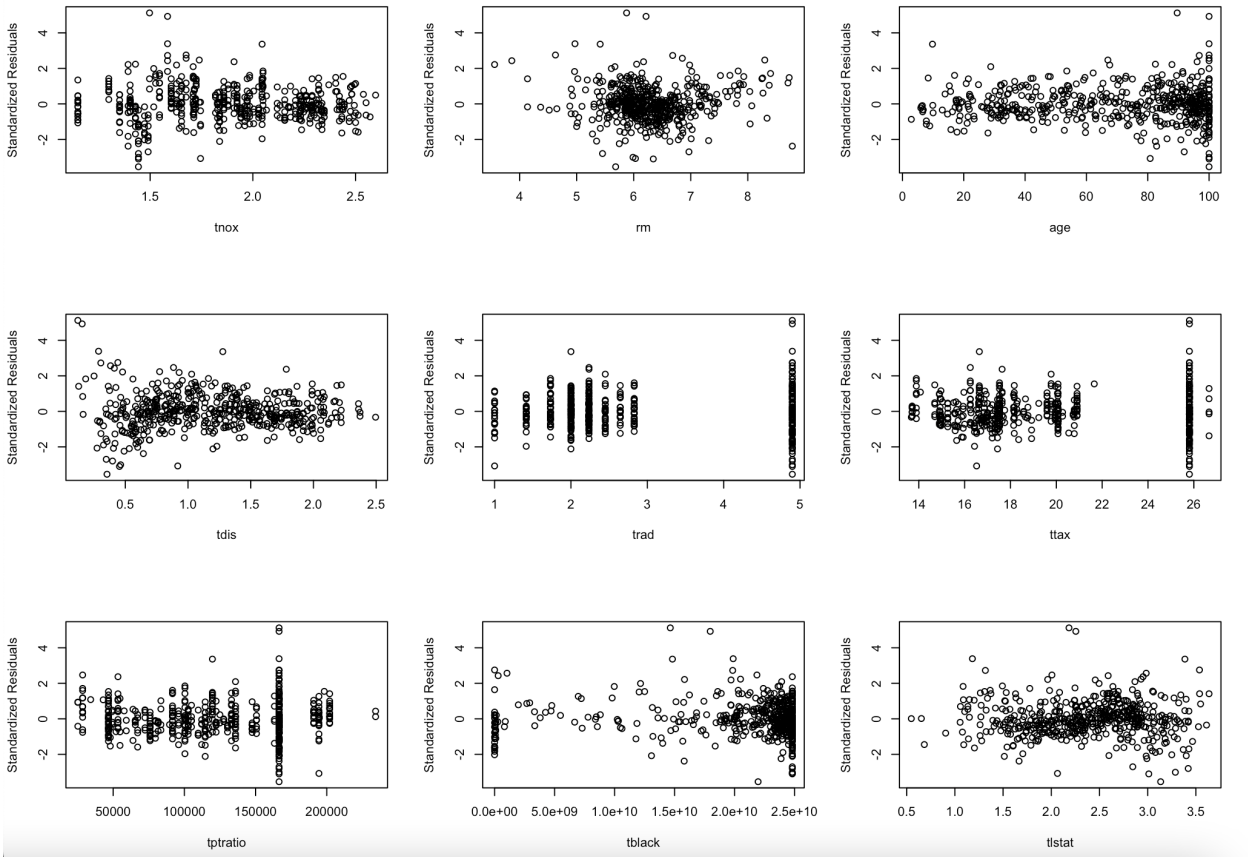

Fig. A.9 Standardized Residual Plots of Final Variables