DATA2002 Assignment 2: Linear Model for Prices (Boston Housing Census of 1970s)

500551998 500521854 480306993 500506541 500517721

1 Introduction

This report entails the analysis undertaken to examine the performance of effectively predicting the factors that influence the housing prices recorded of the Boston Mass of 1970. While the dataset is dated and may not be effective in comparative analysis of the current housing prices of Boston. The following analysis is to understand the underlying patterns and insights that may enable our modelling of the price influencers.

2 Dataset Description

Domain Knowledge - The study in question was designed to improve upon contemporary methods of determining a monetary value for improved air qualities, and provide a quantitative estimate of willingness to pay for air quality improvements. Their hypothesis was whether people desired cleaner air for higher prices in housing. An aim to describe the monetary value assigned to external entities than the housing itself.

U.S. Census Service for housing in 1970 detailed a questionnaire of 30 questions. Assuming they were given to all of Boston mass on a mail-out mail-back system. We consider they used non-probability sampling where resulting sample describes only the respondents. Hence a convenience sampling.

2.1 Variables, Reliabilty and Bias

The dataset consists of a small sample of 506 observations of 14 continuous integer variables. The variables of interest are the crime rate, the number of rooms and distance to Boston centers.

The data may involve selection bias and non-response bias due to the unwillingness to participate, and loss of the survey mail in the transport. Further, participants' response bias may occur due to the private questions in the survey. For example, for the rent of the properties, participants may not report the true amount. Lastly, for the median price variable, it has been censored to a maximum $50k. As only 16 observations are impacted, overall the influence is minor. Otherwise it would be difficult to generalise the predictions/assumptions past the 50k baseline.

The Boston housing dataset was collected by the U.S. Census Service and has been used extensively throughout the literature to benchmark algorithms. It is superior to others because it contains a large number of neighborhood variables and more reliable air pollution data. Given it is used as a predictive model by a lot of people, it's reliable.

2.2 Initial Data Insights

In the correlation matrix, there is a negative correlation such that between crime rate per capita decreases when the house prices. RM and MEDV, have positive correlation with 0.70, where average number of rooms per dwelling increase by price. Similarly with the positive correlation 0.25 between DIS and MEDV, yet correlation is weak.Also we observe evidence for collinearity between RAD and TAX variables which may need to be dropped. (Appendix fg.3)

Our initial inferences being crime having a depreciating linear relationship with price and the distribution of price being normal when plotted were well satisfied. Although prices being capped at 50k are suggestive for inaccuracies.The distribution shows the median value of a home is to be predicted, for example, the median value of 60 owner-occupied homes is $20,000.

Figure 1: Price vs Crime (left) & Price Distribution (right)

3 Model Selection & Analysis

3.1 Data Transformations

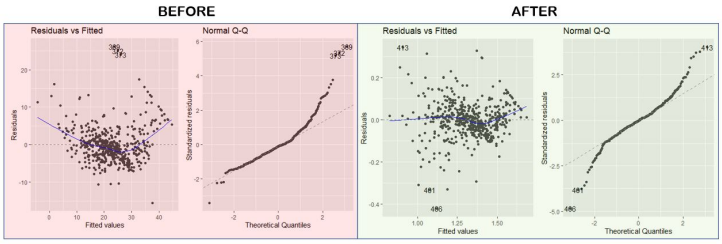
From the initial data insight, some variables do not have clear linear relationships with the median price of the properties (See Figure 2). Therefore, some transformations are applied. Firstly, the MEDV , CRIM , INDUS and DIS have been transformed into log base 10 scales. Meanwhile, the variable B has transformed into the scale of power of three. Further, the ZN variable is scaled based on the number of observations in the dataset and to decimal format. From the plot, an improved linear relationship between these five covariates with the median price can be observed after these transformations. Additionally, as CHAS variable is binary data, it has transformed from numeric into factor.

3.2 Assumptions

Both the pre and post transformation data have satisfied the assumption of independence, as census is independent for the vast majority of responses (refer to Figure 3).

Before transformation, the data violated many assumptions, most critical of which is the linearity. Homo-skedasticity has been somewhat violated, as particularly at the extremes, data points tend to be more spread out and above, rather than below the curve. The residuals have a positive skew, and hence have violated the normality assumption (Appendix fg.4).

The post-transformation performs better; first, linearity is improved and the residuals seem to have an approximately constant variance, with some more spread visible to the left of the plot. However, this is not substantial and as such the homoskedasticity assumption has been met. Whilst the resulting qq-plot is now symmetric, it reveals a larger kurtosis than what is found in a normal distribution, which may lead to compromised inferences and larger or narrower confidence and prediction intervals.



3.3 Choosing Best model

RAD and TAX are highly correlated, which may lower model power and therefore dropped. A backward search using AIC was then implemented to test the effect of the other variables. The backward method was chosen as it doesn't produce suppressor effects. The variables INDUS, ZN and AGE are not significant in the model (no effect of MEDV) with p-values >0.05. Further checking is done through an exhaustive approach which resulted at 6 variables to be the same as the optimal model. While automated tests give optimal models, checking relevance is essential. Variables such as NOX, DIS, PTRATIO and CHAS make a change on the quality of life which explains their effect on house prices. For the RM variable, a larger number of rooms means a bigger more expensive house. Lastly, B and LSTAT variables directly relate to people living in less expensive houses.

4 Hypothesis & Results

We use out of sample method rather than in-sample to prevent overfitting and observe predictability against new data samples. Testing the inference crime and price of households having strong linear dependency resulted in similar error rates and R^2 squared values (in appendix). However, the effectiveness of the crime parameter was negligible as it required additional data at a small improvement of the model.With relatively low error rates and from the R^2 value more than 75% of the variability of prices is explained by the regression of the variables.Also with the significant p-value we can assume the model effectively predicts the price of the boston houses with minor variation. With comparison of the full and reduced model our testing suggests that we conclude that at least one independent variable of the final model is effective in explaining the variability explanatory or predictive power.

4.1 Final Model - Assumptions

The resulting model's residual vs fitted graph shows that as majority of the points are spread out consistently with no obvious fanning, the homoskedasticity and normality assumptions are satisfied (Appendix fg.5). The linearity is well satisfied as the majority of the points are close to the diagonal with minimal curvature at the extreme. Although with the sample size of over 500 the CLT should be in effect.

5 Discussion - Parameter Estimates

Parameter Estimates suggest that on average, a one unit increase in nitric oxide rates gives -40.09% change in price, with the number of rooms and river body providing 4.3% and 5.9% increase respectively. While other parameters provide minor contributions (Figure 2). Although the proportion of dark-coloured individuals (B cubed) was found significant due its small coefficient, its inclusion had a negligible impact on RMSE, R^2, MAE.Hence, it was discarded (Appendix fg.6).

Parameters	Coefficients	Influence (1 unit/ %)
Intercept	1.9236	1.924
CHAS1	0.0627	6.300
NOX	-0.4484	-44.800
RM	0.0406	4.100
DIS_LOG	-0.1940	-0.194
PTRATIO	-0.0179	-1.800
LSTAT	-0.0147	-1.500
Observations	506	
R^2	0.747	
Adjusted R^2	0.744	
Residual Std. Error	0.090 (df = 499)	
F Statistic	245.204*** (df = 6; 499)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 2: Final Model Summary

6 Conclusion

In summary it seems the model presents the disfavoring of nitric oxide emission near housing with a strong preference for rooms and environmental factors. While other obvious influences such as crime and tax make a negligible impact on the prediction of the price.

$$log(MEDV) = \alpha + 0.06(CHAS_1) - 0.45(NOX) - 0.04(RM)$$
$$-0.19log(DIS) - 0.02(PTRATIO) - 0.01(LSTAT) + \epsilon$$

7 References

Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Gauthier, J. H. S. (n.d.). 1970 (Housing) - History - U.S. Census Bureau. US Census Service. https://www.census.gov/history/www/through_the_decades/index_of_questions/1970_housing

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing 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

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

Wickham et al (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

8 Appendix

Git Repository Link for Codebase for this Reproducible Report - https://github.sydney.edu.au/rtal2306/LAB-03-CC_early_7.git

Formal Hypothesis Test

- $H_0 : \beta_1 = \beta_2 ... \beta_6 = 0$ vs H_1 : at least one $\beta_i \neq 0$ where i in range 1 to 6
- Resulting F-statistic: $\frac{(SSR_{Full} - SSR_{Reduced})}{MSE_{Full}} = 138.14$
- Resulting P-value < 0.001 with degrees of freedom: 499
- Resulting Adjusted R^2 value: 0.75
- Verdict : Reject H_0

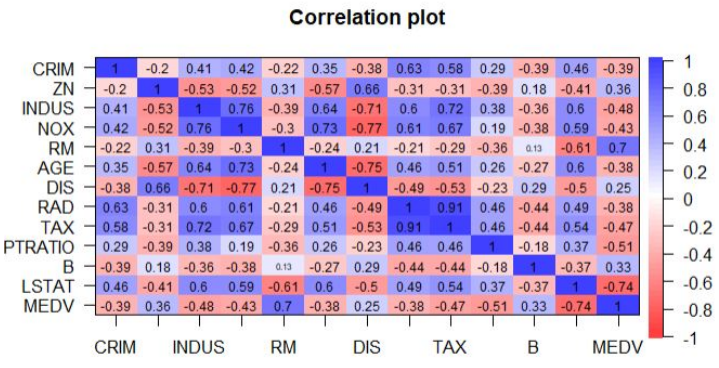


Figure 3: Correlation Matrix

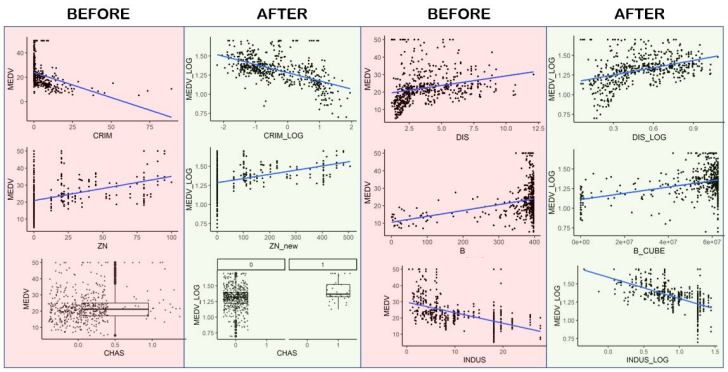


Figure 4: Data transformations - Linearity

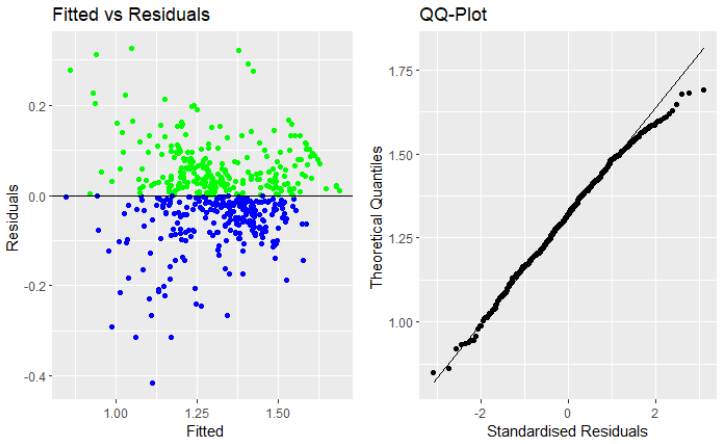


Figure 5: Assumptions Met for Final Model

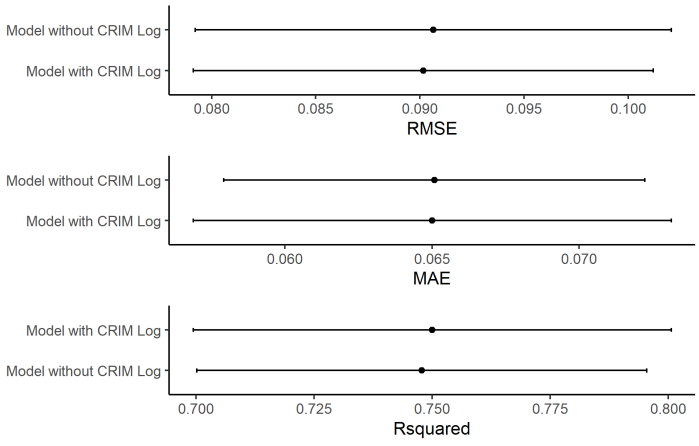


Figure 6: Out of Sample Results