Exploring the Most Significant Features in Predicting House Price

STAT 306 Final Report

Dec 2, 2020

Jason Shao 50963321
Mariana Chen 45839651
Yubo Wang 48684161
Zongyang Gao 95244521

Introduction

Real estate investment accounts for a large proportion of all investment activities. In the investment of real estate, the assessment for the value of the real estate may be affected by a few conditions, such as the quantity of the house, the quality of the house and the environment of the house. In this study, the emphasis will be put on the environment of the house. In the previous research, the author pointed out that the reduction in air pollutant concentrations can lead to the increasing housing price, which indicates the level of natural environment, as well as the quantity and the quality of housing and other neighborhood characteristics, were considered to be the fundamental reasons in making individuals' housing choices (Jr et al.,1978). Since air pollution, one of the natural environments, has been proved to have an impact on the assessment of real estate property, other neighborhood characteristics may also have an impact on the assessment, such as crime rate by town, pupil-teacher ratio by town, and full-value property-tax rate per $10,000. Although large amounts of interest have been paid in the house price modeling, existing models such as the hedonic pricing model proposed by (Jr et al., 1978) focused mainly on the relationship between the level of air pollution and house price, which may not describe the other aspects of factors influencing the house price. Therefore, this project is designed to explore some of the most significant factors on the house price in the area of Boston.

Analysis

Data Collection

Note: we have changed our data set following the proposal submission as we found that were not statistically significant linear relationship between our interested variables.

The data set for our study is *Boston* from the R package MASS. This data set was obtained from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970.

The variables measured are described in Table1. We will use capital letters in the following analysis for consistent.

Variable	**Description**
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 areas per town
chas	Charles river dummy variable (1 Id tract bounds river, 0 otherwise)
nox	Nitrogen oxide concentration (parts per 10 million)
rm	Average number of rooms per dwelling
age	Proportion of owner-occupied units built prior to 1940
dis	Weighted mean of distance 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
black	$1000(\mathrm{Bk}-0.63)^2$ where Bk is the proportion of blacks by town
lstat	Lower status of the population (percent)
medv	Median value of owner-occupied homes in $1000s.

Table1. Variables recorded in the study

Statistical Modelling

We decided to perform model selection and evaluate the subset regression models using different types of criteria, including coefficient of multiple determination (R^2), adjusted R^2 and Mallows' Cp. Performing a thorough analysis of this model allows determining the "best" regression model indicating that the most significant explanatory variables for house pricing.

Results

Illustrated in Fugure1 and Figrue2, the values of our original response variable MEDV are approximately distributed normally despite some outliers and the data seem to be capped at 50. Therefore, we decided to remove the values greater than 50, and use the new data set as our further analysis.

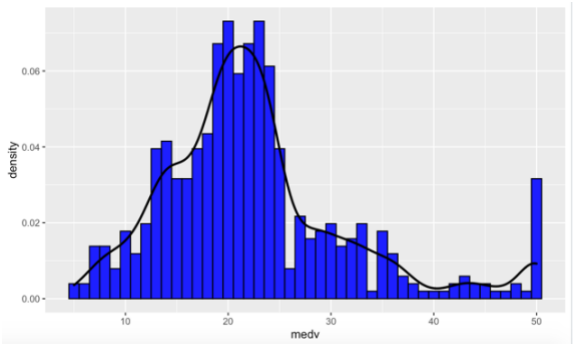

Figure1. Histogram of Original MEDV Values.

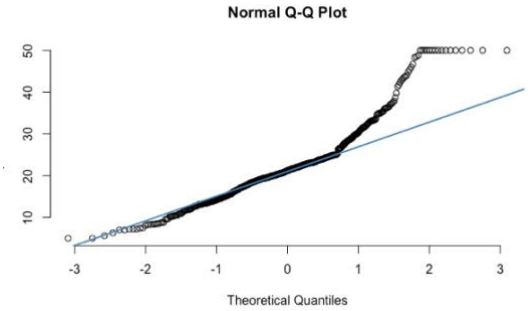

Figure2. Normal Q-Q Plot of Original MEDV Values.

As can be seen from Figure3, rm and medv are positively correlated (0.69) while LSTAT and MEDV are negatively correlated (-0.76). Based on these two observations, we created scatterplots of MEDV varying with LSTAT and RM respectively. Clearly, Figure 4 and Figure 5 show that MEDV increases as RM increase linearly (MEDV = -30.0051 + 8.2686*RM), and the MEDV decreases as LSTAT decreases (MEDV = 32.54041 – 0.84374 * LSTAT) though it does not seem to follow an exactly linear pattern. In addition, it's noticeable that INDUS and DIS have a high correlation of -0.71, NOX and DIS have a high correlation of -0.77, AGE and DIS have a high correlation of -0.74. These variables will be removed in further reduced models to reduce collinearity.

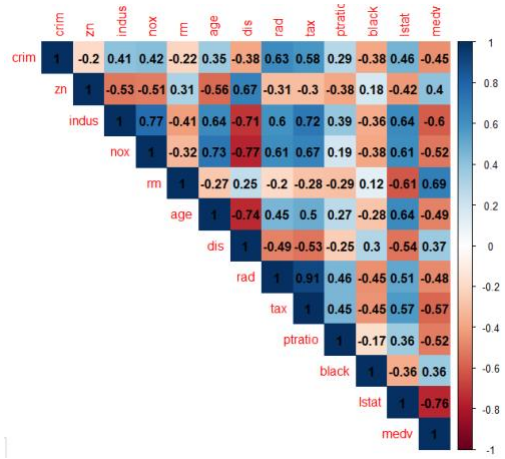

Figure3. Correlation Plot Between the Variables

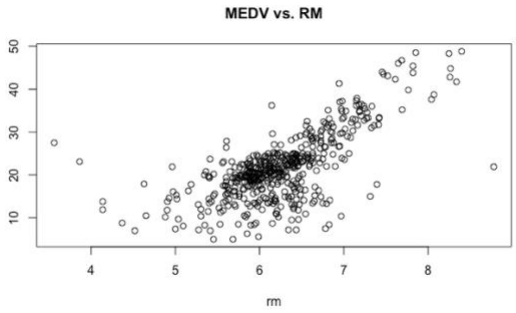

Figure4: Scatterplot of MEDV vs. RM.

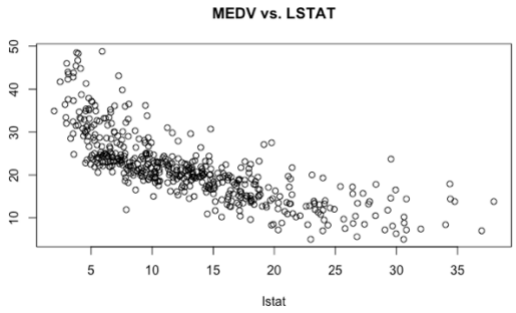

Figure5: Scatterplot of MEDV vs. LSTAT

In order to assess the most significant variables, we started with the full model including all the explanatory variables. Then, we generated different reduced models by adding or removing some explanatory variables that are highly correlated. The linear equation of the full model is depicted in Table 2 for Eq (1) with the largest R^2. Residual plot of full model in Figure 3 suggests that there is equal variance across data despite some outliers, so a linear model would be applicable. INDUS and CHAS were removed since their p-values were not significant, the re-fitted linear equation is described in Table 2 for Eq (2). Besides, AGE, NOX, INDUS and DIS were removed to prevent collinearity and the re-fitted linear equation is expressed as Eq (3) in Table 2.

	Regression Equations	$\mathbf{R^2}$	$\mathbf{adjR^2}$
(1)	MEDV = 32.23 – 0.11Crim + 0.04Zn – 0.04Indus + 0.45 Chas1 -12.4Nox + 3.76 Rm – 0.02Age – 1.21Dis + 0.25 Rad – 0.01Tax – 0.84 Ptratio + 0.01 Black – 0.35 Lstat	0.7777	0.7716
(2)	MEDV = 32.46 -0.11Crim + 0.04 Zn -13.01 Nox+ 3.80Rm – 0.02Age – 1.18Dis + 0.27Rad – 0.02Tax – 0.86Ptratio + 0.01Black – 0.35Lstat	0.7772	0.772
(3)	MEDV = 17.6 – 0.08Crim + 0.01Zn + 0.33Chas1 + 4.04Rm + 0.24Rad – 0.01Tax -0.82Ptratio + 0.01Black – 0.38Lstat	0.7487	0.744

Table2. Models generated, corresponding R^2 and $adjR^2$.

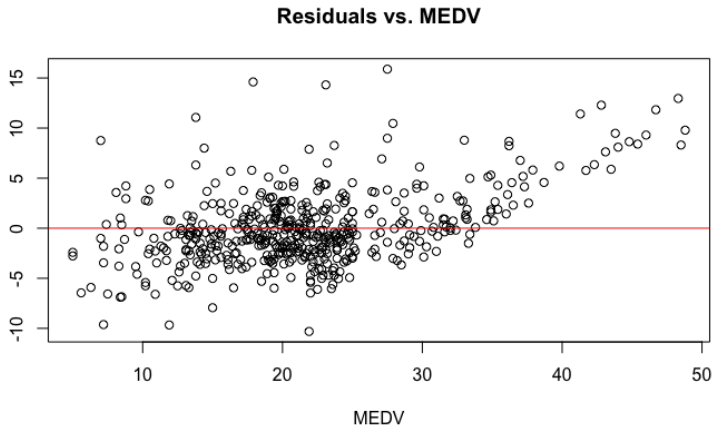

Figure 3: Plot of residuals vs. MEDV

However, according to ${R_p}^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res(p)}}{SS_T}$, full model will always yield largest ${R_p}^2$.

Thus, we cannot find an "optimum" value of ${R_p}^2$ for a sublet regression model. The information concerns Mallows' Cp statistic and p in Table 3 and Figure 4 suggests that model (2) with 11 variables is the best as Cp is closest to p.

No. of Variables	Cp	adjR^2
1	418.82817	0.5764869
2	251.37636	0.6550936
3	133.24859	0.7108422
4	99.95332	0.7268685
5	73.76467	0.7396072
6	56.39654	0.7482273
7	48.32193	0.7524792
8	35.28918	0.7591032
9	24.01337	0.7649187
10	13.89357	0.7702073
11	11.07839	0.7720275
12	12.37174	0.7718880
13	14.00000	0.7715872

Table 3. Mallows' Cp statistic and adjR^2

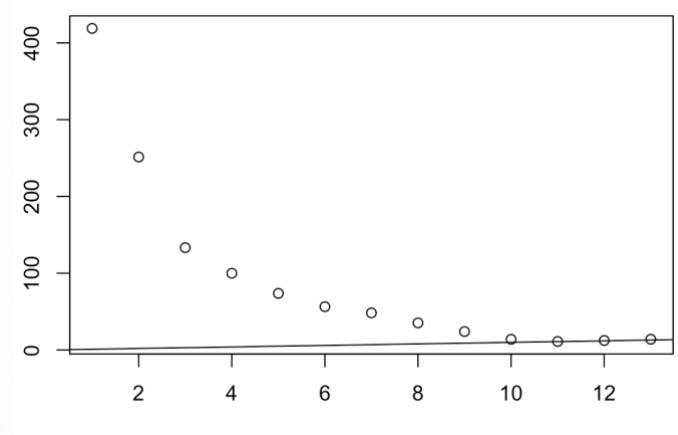

Figure 4: Plot of Mallows' Cp statistic and p

Table 5 which illustrates the output of our best model (model 2) leads us to the conclusion that LSTAT, RM and PTRATIO are the most significant variables in predicting the house price in Boston.

- While taking all other variables constant, for every 1 unit increase in the number of rooms the house price will increase by approximately $3800. It's intuitive that more rooms require more space, it will apparently cost more.
- While taking all other variables constant, the house price will decrease by approximately $354 as the proportion of lower status population increase by 1%. Given an area with more "lower class" citizens, if the price of housing is extremely high, it will make them unaffordable.
- While taking all other variables constant, the house price will drop by approximately $857 as the pupil-teacher ratio increase by 1 unit.

Since MEDV was not exactly linearly related to LSTAT, a transformation of LSTAT was included in the full model. Transformed models with their corresponding $adjR^2$ and R^2 are displayed in Table 4. It's observable that the model including log (LSTAT) will yield the highest adjR^2.

Models	$adjR^2$	R^2
lm (medv~. + I(lstat^2), data = dd)	0.8022	0.8079
lm (medv~. + log(lstat), data = dd)	0.8079	0.8134
lm (medv~. −indus−chas + I(lstat^2), data = dd)	0.8027	0.8075
lm (medv~. -$i$ndus−chas+ log(lstat), data = dd)	0.8083	0.813

Table 4. Models with Transformed Variables

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	32.464301	4.112776	7.894	2.02e-14
crim	- 0.106461	0.026012	-4.093	5.00e-05
cn	0.036202	0.011207	3.230	0.001321
nox	- 13.009570	2.923825	- 4.450	1.07e-05
rm	3.796139	0.355095	10.690	< 2e-16
age	- 0.023363	0.010637	- 2.196	0.028538
dis	- 1.182636	0.156782	- 7.543	2.33e-13
rad	0.265798	0.051050	5.207	2.86e-07
tax	- 0.015051	0.002707	- 5.560	4.49e-08
ptratio	- 0.856513	0.103764	- 8.254	1.50e-15
black	0.007964	0.002126	3.746	0.000201
lstat	- 0.353550	0.042276	- 8.363	6.75e-16

Table 5. R output of model 2

Conclusion

We observed the high significance of ‘RM’, ‘PTRATIO’, ‘LSTAT’ in our model, but the importance of ‘NOX’ is not significantly different from other variables. The reason is that the air quality will not be significantly different in different areas of the city and will affect each other due to gas mobility. Another important factor is the ‘pupil-teacher ratio by town’. Usually, a lower pupil-teacher ratio means better education, which also means higher income and higher investments in the next generation of a family. Thus, ‘PTRATIO’ can indicate the house price indirectly. From the model comparison using different evaluation indicators, we found that model using MEDV as response variable, all variables plus log of LSTAT, exclude INDUS, CHAS as explanatory variables gives the best result with highest adjusted R^2., meanwhile being a relatively simple model. However, if we fit a model with fewer explanatory variables (such as 4) can still give us a well-performed adjusted R^2. Thus, we should consider the trade-off between model complexity and prediction performance of models in our further research. Some other

transformations to data can also be applied such as polynomials transformation. Also, the data is collected half a century ago, it might not be suitable to make predictions for data collected in recent years. In conclusion, compared to air quality, the factors affecting the house price in an area are the level of education (indicated by 'Pupil-teacher ratio by town'), house area (indicated by 'Average number of rooms per dwelling') and income level of the residents (indicated by a reverse of 'Lower status of the population').

Reference

Jr, D. H., & 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