Median value of owner in Boston House price data set

LAB-01-RE_early_7[a]

[a]The University of Sydney, NSW, 2008

This manuscript was compiled on November 11, 2021

This report is devised to study the Boston housing dataset collected by the U.S Census Service and investigate the correlation between the variable "MEDV" and other attributes in the dataset. In approaching this task, the research questions are subdivided into several aspects to provide a comprehensive illustration. The report starts with an IDA and model selection. Assumption checking and discussion of the result are included as the later steps. With the use of multiple regression, we have been able to generate the model of 'MEDV' and achieved an adjusted R-squared of 0.78 and root-mean-square error of 0.19 as our formal findings.

Introduction

Background. In the first phase, an initial exploration of the attribute "MEDV" and its relationships with other associated variables would give us an overview of the dataset. As we progress through several experiments with modeling, a simple and easily interpretable "MEDV" model with less insignificant variables has been produced by stepwise regression. The model selection step has provided us with insight into intricate correlations between "MEDV" and other variables, allowing the model to be improved to a statistically significant extent.

Meanwhile, we have developed an interest in studying the key assumptions underlying our multiple regression model. The assumptions including linearity, independence, homoscedasticity and normality have all been scrupulously examined to ensure our model is valid and reliable. The simplified model has provided sufficient information to predict how a proportion of change in each of the other variables has reflected on our target variable "MEDV" while holding the remaining variables constant as the answer to our main research questions towards how different variables interact with "MEDV." In the concluding step, we evaluate several performance indicators of the model to analyse the model's efficiency.

Data Set. The origin of the Boston housing data is Natural, this publicly available dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. There are 14 columns and 506 rows in this data set, and all of the attributes are numerical. All of the variables are towards describing a Boston suburb or town. The first analysis of the dataset was made by David Harrison Jr. and Daniel L. Rubinfeld, called Hedonic housing prices and the demand for clean air. . . (2)

Our target variable "MEDV" describes the Median value of owner-occupied homes in $1000's and a Detailed dictionary of the data description of all the variables will be attached to the end of this report.

During the study of the dataset, It is notable that the prices of homes are capped at 50, which was caused by censorship of US census service, restricting the range of price to 0-50(k). This leads to the prices of any houses whose value were more than 50k to be unknown and studies have proven there have been houses that were valued more than 50k(3). Hence our study of the 'MEDV' variable is restricted due to the ceiling of '50.'

Moreover, some of the house values have been incorrectly recorded, this is also evidenced by the study of Otis W. Gilley(3) which also limits the accuracy of our 'MEDV' model.

Analysis

Assumptions. Assumption checking was carried out before and after performing variable selection to guarantee validity. Residuals are symmetrically distributed above and below zero, thus the linear assumption is reasonable. The house from Boston is not related to one another, which follows the Independence assumption. There are some outliers in the Residuals vs Fitted graph, therefore we use the central limit theorem. The residuals do not appear to be fanning out or changing their variability over the range of the fitted values so the homoskedasticity assumption is met. Residuals QQ plot shows that most points are close to the straight line, also relying on the central limit theorem, the normality assumption is satisfying.

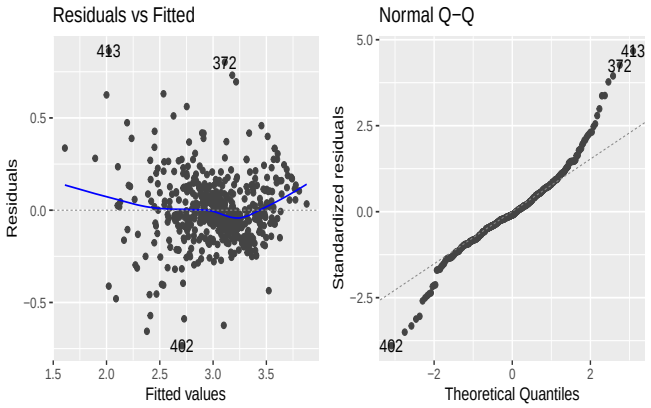


Fig. 1. Residuals plots for the final regression model

Fact

Predicting the median value of a home is complicated because it is related to many factors. Our final model tells us that the relationship between the nitrous oxide level and the median value of a home is very close

Model Selection. Log transformations were performed in attempt to fit the `MEDV` variable into a linear relationship with the dependent variable.

Firstly, we fit the full model, although the value of r squared shows that 74% of MEDV variables are explained by this regression model, there are some variables that have high p-values(see pic 1) and so are not significant, so we need to drop them and find a more accurate one. We then use backward and forward search using AIC to find out 2 regression models and then compare them and choose the one with smaller AIC value which will be the more appropriate one. The models found by backward and forward search are quite similar shown in pic2. They have the same r-squared(0.7406) and RMSE (4.736) value, and so the final model is: $\hat{MEDV} = 36.34 - 0.11 \times CRIM + 0.05 \times ZN + 2.72 \times CHAS - 17.38 \times NOX + 3.8 \times RM - 1.49 \times DIS + 0.3 \times RAD - 0.95 \times PTRATIO - 0.52 \times LSTAT - 0.01TAX + 0.01B$ However, we found that the residuals vs fitted plot for this model are not so well (pic 4), the blue line appears to be quite curved, so we change the y variable to the log(y) and get a new model (pic 5). This model has a higher r-squared value which means more variables are explained by it and a lower RMSE value, and all x variables are significant. Thus, we decide to take this model as the final model. $\hat{LMEDV} = 4.08 - 0.01 \times CRIM + 0.001 \times ZN + 0.1 \times CHAS - 0.72 \times NOX + 0.09 \times RM - 0.05 \times DIS + 0.01 \times RAD - 0.04 \times PTRATIO + 0.0004 \times B - 0.03 \times LSTAT - 0.0006 \times TAX$

Results

Inferences.

$$\begin{aligned} \text{LMEDV} = & 4.08 - 0.01(\text{CRIM}) + 0.001(\text{ZN}) + 0.1(\text{CHAS}) \\ & - 0.72(\text{NOX}) + 0.09(\text{RM}) - 0.05(\text{DIS}) \\ & + 0.01(\text{RAD}) - 0.04(\text{PTRATIO}) + 0.0004(\text{B}) \\ & - 0.03(\text{LSTAT}) - 0.0006(\text{TAX}) \end{aligned}$$

The relationship between the nitrous oxide level and the median value of a home is very close.

Performance. We compared our final model with the full model which contains all variables in the data set being used as predictors. Out of sample performances are tested using "Caret" package at a 10-fold cross-validation.

Our final model formula consists of 11 variables. With the root-mean-square(RMS) , R-squared and mean Absolute Error we are able to measure the data of performance. Meanwhile, by contrasting with the simple model which possesses a RMSE of 4.78, the RMSE of the final model(0.19) is significantly smaller. The difference indicates a smaller prediction error of the final model. In addition, the final model has a larger R-squared value(0.78), that is, on the scale of 0-100% the strength of the relationship between the model "MEDV" and the dependent variable is 78%. We also find that the mean absolute error(MAE) of the final model is 0.14. It is also much smaller than the MAE of the sample model(3.37). Hence, we conclude our final model is the model of good fit to predict the variable 'MEDV".

Table 1. Performance results of models

Attributes	Full Model	Final Model
R-squared	0.9427	0.7787411
In Sample RMSE	2.2533	0.1915
In Sample MAE	1.5293	0.1374
Out of Sample RMSE	2.2870	0.1932
Out of Sample MAE	1.5336	0.1389
Out of Sample r-squared	0.9378	0.7853

Discussion

Limitations. Apart from the limitation of the data source discussed in the data description. The theoretical limitation we have encountered arose from the principal drawbacks of stepwise multiple regression. Studies(1) have indicated that bias may exist in the process of parameter estimation and inconsistencies of model selection algorithms can also become problematic. Relying on a single best model can also bring risk factors.

Conclusion. Through the assessment of various performance indicators including R-squared and the root mean square error of 'MEDV' model, our report concludes that the model is suitable for the purpose of the goodness of fit. However, a series of limitations that have been outlined in this report should also be contemplated to ensure the integrity and accuracy of the model.

Indeed, multiple regression analysis is powerful in modelling certain variables in the Boston housing dataset. By applying a backward/forward search, we have reduced the number of variables to find the subset of variables within our dataset in the best performing model. However, other prediction methods may create a better model than multiple regression. It is through consistent experimentations with various algorithms such as k-nearest neighbours and polynomial regression that we can discover the most appropriate model that fits the variable and develop our comprehensive understandings through our collaborative endeavour to develop and innovate.

References

GitHub repository

(1)Mark J. Whittingham1 , Philip A. Stephens2 , Richard B. Bradbury3 & Robert P. Freckleton4 . Why do we still use stepwise modelling in ecology and behaviour? [Avilable at:https://eprints.ncl.ac.uk/file_store/production/56364/AE9762E1-23E2-4C9D-9518-5BCDAD47FAAB.pdf]

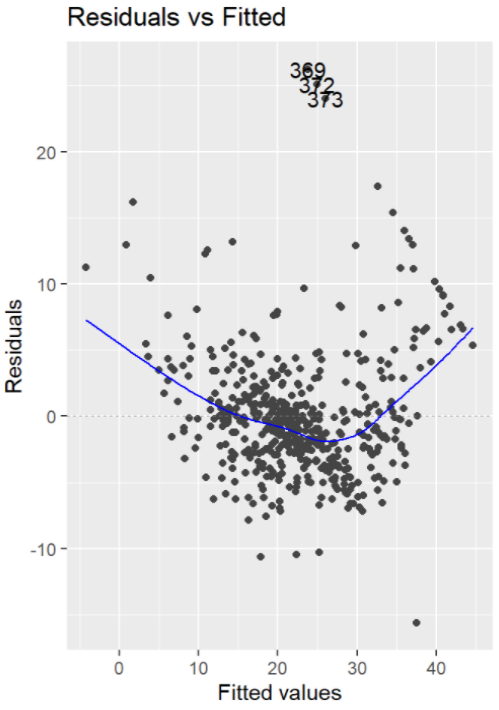
(2)Harrison, David & Rubinfeld, Daniel, Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management (1978), Journal of Environmental Economics and Management

(3)O. W. Gilley, On the Harrison and Rubinfeld Data (1996), Journal of Environmental Economics and Management

Appendix

```
##
## Call:
## lm(formula = MEDV ~ ., data = hp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
## ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
## INDUS        2.056e-02  6.150e-02   0.334 0.738288
## CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
## NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## AGE          6.922e-04  1.321e-02   0.052 0.958229
## DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
## PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## B            9.312e-03  2.686e-03   3.467 0.000573 ***
## LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

pic 1



pic 4

```
## Call:
## lm(formula = LMEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + CHAS +
##     B + ZN + CRIM + RAD + TAX, data = hp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73400 -0.09460 -0.01771  0.09782  0.86290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0836823  0.2030491  20.112  < 2e-16 ***
## LSTAT       -0.0286039  0.0019002 -15.053  < 2e-16 ***
## RM           0.0906728  0.0162807   5.569 4.20e-08 ***
## PTRATIO     -0.0374259  0.0051715  -7.237 1.77e-12 ***
## DIS         -0.0517059  0.0074420  -6.948 1.18e-11 ***
## NOX         -0.7217440  0.1416535  -5.095 4.97e-07 ***
## CHAS         0.1051484  0.0342285   3.072 0.002244 **
## B            0.0004127  0.0001071   3.852 0.000133 ***
## ZN           0.0010874  0.0005418   2.007 0.045308 *
## CRIM        -0.0103187  0.0013134  -7.856 2.49e-14 ***
## RAD          0.0134457  0.0025405   5.293 1.82e-07 ***
## TAX         -0.0005579  0.0001351  -4.129 4.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1898 on 494 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7844
## F-statistic: 168.1 on 11 and 494 DF,  p-value: < 2.2e-16
```

pic 2

	Forward model		Backward model	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	36.34	**<0.001**	36.34	**<0.001**
LSTAT	-0.52	**<0.001**	-0.52	**<0.001**
RM	3.80	**<0.001**	3.80	**<0.001**
PTRATIO	-0.95	**<0.001**	-0.95	**<0.001**
DIS	-1.49	**<0.001**	-1.49	**<0.001**
NOX	-17.38	**<0.001**	-17.38	**<0.001**
CHAS	2.72	**0.002**	2.72	**0.002**
B	0.01	**0.001**	0.01	**0.001**
ZN	0.05	**0.001**	0.05	**0.001**
CRIM	-0.11	**0.001**	-0.11	**0.001**
RAD	0.30	**<0.001**	0.30	**<0.001**
TAX	-0.01	**0.001**	-0.01	**0.001**
Observations	506		506	
R^2 / R^2 adjusted	0.741 / 0.735		0.741 / 0.735	
AIC	3023.726		3023.726	

pic 5