Exectutive Summary | Boston Data Set

480440172, 490175477, 480561916, 480550367, 470500790

DATA2002 Group M13-07

The University of Sydney 2019 Semester Two

A city's housing is a crucial and important factor for the well-being of residents and plays quite a significant role in the sustainability of an economy. Rising house prices generally equate to higher rates of economic growth [2] due to encourage consumer spending. But this can also see first-time buyers and families struggling with increasingly unaffordable prices in the market. This leads us to the purpose of this investigation, which is to determine and discover what factors and variables may be significantly impacting the median house prices. The findings of this report suggest that factors such as crime rates, nitric oxide concentration and several other variables do have some statistically significant effect on the median house price in boston.

Introduction

The main purpose of this report is to explore and evaluate the factors affecting Boston's housing prices by using multiple regression analysis. Once a model is created and meets the assumptions, we then ask the interesting questions to determine whether nitrious oxide concentration and distances to Boston employment centres significantly contribute to influencing the median value of owner-occupied homes. Housing prices can be affected by factors such as but not limited to: location, demographics, unemployment rates, number of rooms, etc [2]. Multiple regression uses multiple predictor variables to predict an outcome which justifies why this model can be used for the Boston Housing dataset analysis and is therefore suitable for this analysis.

The Data set

To complete this investigation a dataset containing information collected by the U.S Census Service in 1978 is used. This sample consists of 506 entries and holds information concerning housing in the area of Boston Mass. The dataset was obtained from StatLib archive but was originally published by Harrison, D. and Rubinfeld, D.L.. The CSV file used in this report was obtained from the StatLib archive [3] and has a total of 506 entries.

Additionally, the dataset has 14 variables which represent; crime rate per capita, the proportion of residential land zones for lots over 25,000 square feet, the proportion of non-retail business acres, if it was bound by a river, nitric oxides concentration, the average number of rooms per dwelling, proportion of owner occupied units since 1940, the distances to five Boston employment centres, the accessibility of radial highways, the pupil teacher ratio by town, the proportion of African Americans by town, the percentage lower status of the population and lastly, the median value of owner-occupied homes in \$1000's.

Analysis

Model Selection and Assumptions. After initial analysis of the dataset, CHAS is dropped from the data frame, because it is a binary dummy variable that states whether a track bounds a river, and since it is binary, it can not be used in multiple regression. The variable MEDV is the Median value of owner-occupied homes in \$1000's and is the variable of interest since it will assist in establishing what affects the prices of homes. We perform a multiple regression analysis on the variable MEDV using a backwards stepwise procedure on a full model (Apendix). To rid of insignificant variables we use the AIC backwards search model. After the step wise procedure, the variables AGE and INDUS were dropped as they were proven to be statistically insignificant to MEDV in our model.

We then decided to check the model and our assumptions, and after viewing our residuals versus fitted plot, we decided to do a log transformation on the model since there was a distinct pattern. Our regression assumptions were met after our log transformation:

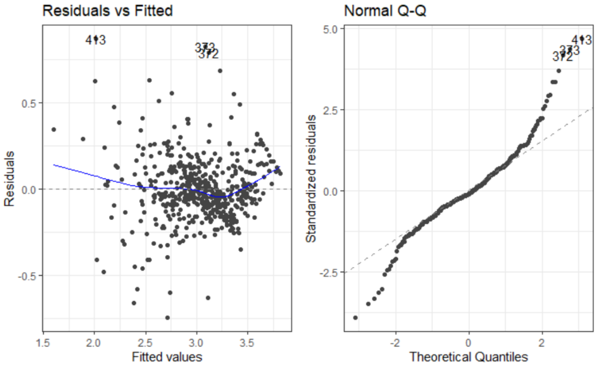


Figure 1: Model Assumptions

As seen in the residuals versus fitted plot there is no obvious patterns and therefore it does not seem that the model has been misidentified. Regarding homoskedasticity, the residuals don't appear to be fanning out or changing their variability over the range of the fitted values so the constant error variance assumption is met. in the QQ plot, the points are reasonably close to the diagonal line, although it is possible to argue that the points may be slightly over-dispersed. Considering that the sample data initially contained 506 entries it is understandable to have some deviation which is not severe enough to cause too much concern. The normality assumption is at least approximately satisfied. Looking at the R^2 value (multiple R-squared) from the summary output (Appendix), ~79% of the variability is explained by the regression on the variables listed above. The fitted model is seen below:

$$\widehat{MEDV} = 4.094 - 0.101 \times \text{CRIM} + 0.001 \times \text{ZN} - 0.687 \times \text{NOX} + 0.092 \times \text{RM} - 0.053 \times \text{DIS} + 0.014 \times \text{RAD} - 0.001 \times \text{TAX} - 0.039 \times \text{PTRATIO} + 0.001 \times \text{B} - 0.029 \times \text{LSTAT}$$

To further investigate the effect on house prices, we may look at some variables of interest from the model. This well provide some more explanation to our main inquiry and may also prove our model correct.

What effect does DIS have on MEDV?

The weighted distances to five Boston employment centres (DIS) is significant in the model with a p-value of 5.21×10^{-12}. In theory this makes sense, because the closer the employment centres the more expensive houses may be and vice versa. However, we cannot trust the p-value without plotting the data first and determining what level of significance DIS has on MEDV individually.

Hypothesis: $H_0: \ \beta_1 = 0$ vs $H_1: \ \beta_1 \neq 0$
Assumptions:

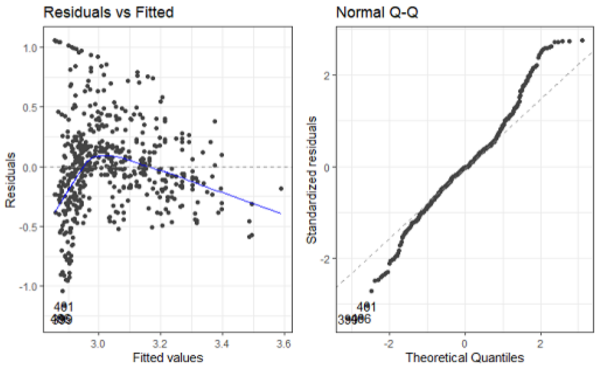


Figure 2: DIS Assumptions

There is no completely obvious pattern in the residual vs fitted values plot, therefore, it does not appear that we have misspecified the model. Regarding the homoskedasticity of the model, the residuals do not appear to be changing their variability over the range of fitted values, and there is no significant fanning out, therefore the constant error variance assumption is satisfied. The QQ plot allows us to assume normality becuase the points are reasonably close to the line. It is important to note that there are many variables in the data set, meaning the QQ plot will not be perfect. The normality assumption is at least approximately satisfied.

Test Statistic: $t_0 = \dfrac{1.0916}{0.1884} = 5.79$

P-value: $2P(t_{504} \geq 5.79) < 0.0001(1.206612 \times 10^{-08})$

Decision: Since the p-value is less than 0.01 we reject the null hypothesis and conclude that DIS does indeed have some significant effect on house pricing. The plot for this model can be seen the Appendix (Figure 6).

What effect does NOX have on MEDV?

Nitric oxides concentration (NOX) may also impact house pricing, since places with higher NOX concentrations will usually have more cars, and therefore be more dense which most likely equate to higher house prices.

Hypothesis: $H_0: \ \beta_1 = 0$ vs $H_1: \ \beta_1 \neq 0$

Assumptions:
All assumptions are met. Autoplot can be seen in appendix (Figure 3).

Test Statistic: $t_0 = \dfrac{-33.91606}{3.196337} = -10.61091$

P-value: $2P(t_{504} \geq 10.61) < 0.0001(7.065042 \times 10^{-24})$

Decision: Since the p-value is less than 0.01 we reject the null hypothesis and conclude that NOX does indeed have some significant effect on house pricing. The following figure displays this model:

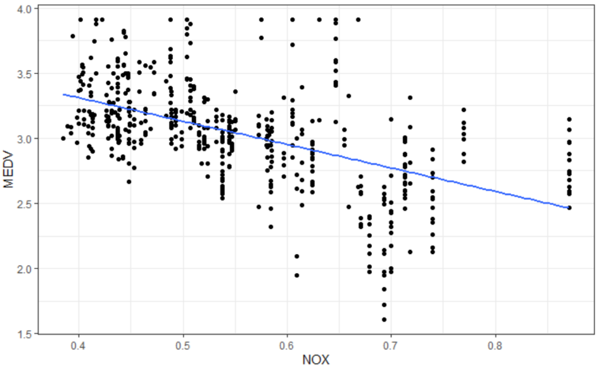


Figure 4: MEDV VS NOX

The results are quite interesting, becuase as NOX increases MEDV decreases, suggesting that areas with higher Nitric oxides concentrations consist of cheaper house than areas with less concentration.

Results

There is evidence that indicates a weak correlation between NOX and Median value of the property and DIS on MEDV. Additionally, when a linear regression is performed for them, a relatively weak linear relationship was found for the comparisons between these 3 variables. This was checked by obtaining the Rˆ2 values of them which were 0.2601 and 0.1157 respectively.

From the multiple regression analysis, there are 2 variables that were found not to be significantly correlated. Namely AGE and INDUS. The remaining 11 variables showed significant correlation with the Median value of the properties.

Discussion and Conclusion

This report generally has many interesting findings regarding the house market. However, despite this there are many limitations to this study which potentially have the ability to alter and influence results. The boston dataset used may have some issues. For instance, some of the data was sourced from outside of Delve and may be somewhat suspect since there is not a definite reliable source for some of the comparisons in the dataset. Being published in 1978, it is clear the data is also quite relatively old. This would not necessarily impact the results but may introduce some bias to the general questions about the median house market in today's society. In a future analysis, it would defenifenty be more appropriate to investigate a data set that is recent and reliable to ensure contemporary and relevant results. In addition to this, there is some evident departure on the QQ plot and there is a small pattern of over-fitting, however, the dataset is quite large and this is somewhat expected when dealing with larger datasets. Despite this, perhaps in future investigations it will be more fitting to have a deeper analysis into the variables and a wider range of model selection.

In conclusion, from our multiple regression, we found that housing prices in Boston is found to be not significantly correlated with (AGE) and proportion of non-retail business acres per town (INDUS). This model was reasonably accurate in predicting median house prices ($R^2 = 0.79$). The linear models show that nitrious oxide levels and distance to employment centres are a relatively weak direct relationship, so we can assume that it does play somewhat of a role in influencing housing price.

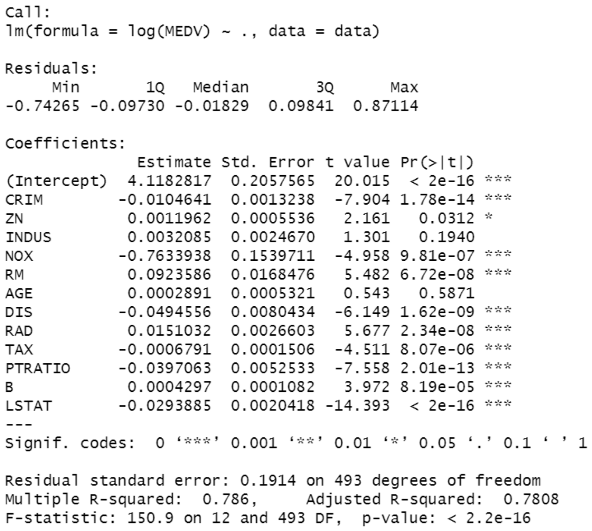
References

[1] Harrison, D, and D.L Rubinfeld. "Hedonic Prices and the Demand for Clean Air." J. Environ. Economics & Management, StatLib archive , 10 Oct. 1996, lib.stat.cmu.edu/datasets/boston

[2] Pettinger, Tejvan. "How the Housing Market Affects the Economy." Economics Help, EconomicsHelp.org, 24 Apr. 2019, www.economicshelp.org/blog/21636/housing/how-the-housing-market-affects-the-economy/.

[3] Data Set Source: http://lib.stat.cmu.edu/datasets/boston

Appendix



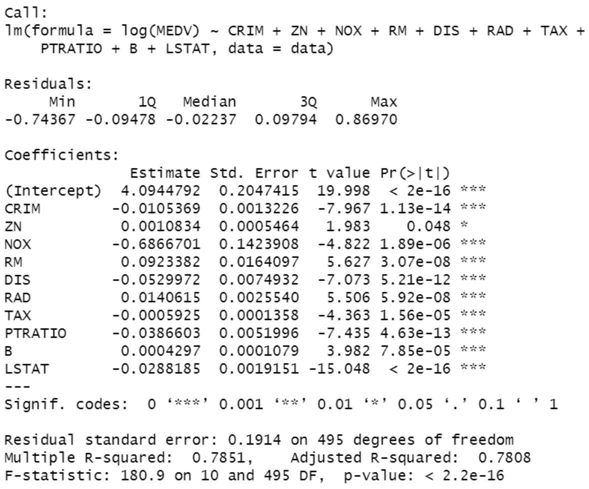
```
Call:
lm(formula = log(MEDV) ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74265 -0.09730 -0.01829  0.09841  0.87114 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.1182817  0.2057565  20.015  < 2e-16 ***
CRIM        -0.0104641  0.0013238  -7.904 1.78e-14 ***
ZN           0.0011962  0.0005536   2.161   0.0312 *  
INDUS        0.0032085  0.0024670   1.301   0.1940    
NOX         -0.7633938  0.1539711  -4.958 9.81e-07 ***
RM           0.0923586  0.0168476   5.482 6.72e-08 ***
AGE          0.0002891  0.0005321   0.543   0.5871    
DIS         -0.0494556  0.0080434  -6.149 1.62e-09 ***
RAD          0.0151032  0.0026603   5.677 2.34e-08 ***
TAX         -0.0006791  0.0001506  -4.511 8.07e-06 ***
PTRATIO     -0.0397063  0.0052533  -7.558 2.01e-13 ***
B            0.0004297  0.0001082   3.972 8.19e-05 ***
LSTAT       -0.0293885  0.0020418 -14.393  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1914 on 493 degrees of freedom
Multiple R-squared:  0.786,	Adjusted R-squared:  0.7808 
F-statistic: 150.9 on 12 and 493 DF,  p-value: < 2.2e-16
```

Figure 7: Full Model



```
Call:
lm(formula = log(MEDV) ~ CRIM + ZN + NOX + RM + DIS + RAD + TAX + 
    PTRATIO + B + LSTAT, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74367 -0.09478 -0.02237  0.09794  0.86970 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.0944792  0.2047415  19.998  < 2e-16 ***
CRIM        -0.0105369  0.0013226  -7.967 1.13e-14 ***
ZN           0.0010834  0.0005464   1.983    0.048 *  
NOX         -0.6866701  0.1423908  -4.822 1.89e-06 ***
RM           0.0923382  0.0164097   5.627 3.07e-08 ***
DIS         -0.0529972  0.0074932  -7.073 5.21e-12 ***
RAD          0.0140615  0.0025540   5.506 5.92e-08 ***
TAX         -0.0005925  0.0001358  -4.363 1.56e-05 ***
PTRATIO     -0.0386603  0.0051996  -7.435 4.63e-13 ***
B            0.0004297  0.0001079   3.982 7.85e-05 ***
LSTAT       -0.0288185  0.0019151 -15.048  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1914 on 495 degrees of freedom
Multiple R-squared:  0.7851,	Adjusted R-squared:  0.7808 
F-statistic: 180.9 on 10 and 495 DF,  p-value: < 2.2e-16
```

Figure 8: Model with dropped AGE and INDUS

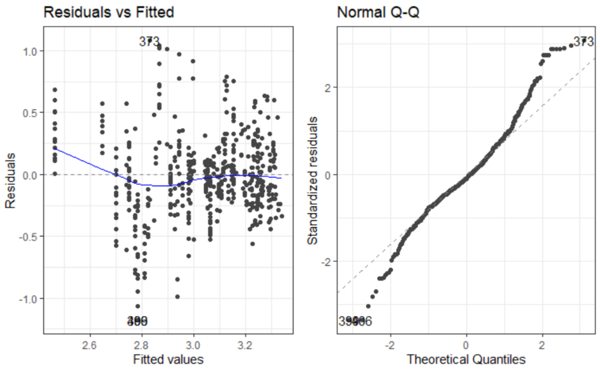


Figure 3: NOX Assumptions

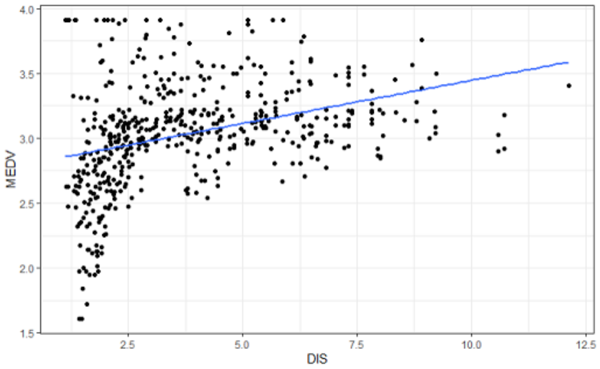


Figure 6: DIS VS MEDV