Understanding House Price in Science

Baiheng Zhou[1], **Seunghyun Lee**[2], **Zijun Shi**[3], **Jun Zhang**[4], **and Yoffie Wu**[5]

[a]The University of Sydney, DATA2002-M13-07

This version was compiled on 11/07/2021

Your abstract will be typeset here, and used by default a visually distinctive font. An abstract should explain to the general reader the major contributions of the article.

House price | Multiple regression | AIC | Nitric oxides concentration | River

Introduction

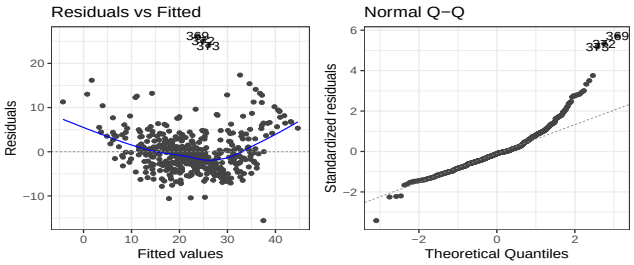
Aim. This report is going to find a way to balance people's wealth and numerous elements that can effect the house price, and predict the price with the dataset by Multiple Regression

The Boston Housing Dataset. There are 506 cases in different towns about Boston housing in 14 variables collected by the U.S Census Service concerning housing in the area of Boston Mass. One of the prototask for this dataset is to predict `MEDV`, the median value of a home. The following list is the metadata.

- 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 acres** per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - **nitric oxides concentration** (parts per 10 million)
- RM - average **number of rooms** per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted **distances** 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
- B - 1000(Bk - 0.63)ˆ2 where Bk is **the proportion of blacks** by town
- LSTAT - % lower status of the **population**
- MEDV - **Median value of owner-occupied homes in $1000's**

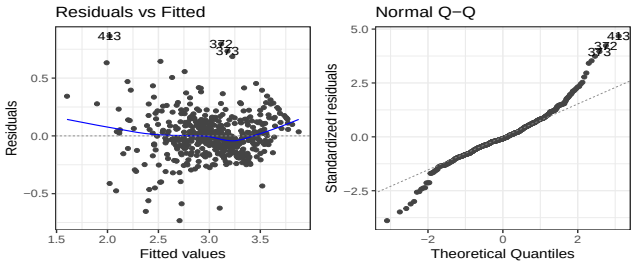
Analysis: Appropriate model selection

Basic model without any transformation.



We first created the multiple regression of Boston house price on the other 13 variables and tested the assumption. We draw residuals versus fitted plots and normal Q-Q plot (fig 1). As residuals vs Fitted values plot shows the non-linear pattern because the residuals are above zero for low house price, then they go below zero and end up again above zero for high house price, we used the log transformation to improve the accuracy.

New model with log transformation.



After log transformation, we now get a Residuals vs fitted plot without obvious pattern (fig 2). Therefore, we were able to satisfy the linearity assumption. Moreover, we met the heteroskedasticity assumption too.As each observation is independent of the other observations, we satisfy the independence assumption. Even though the residuals appear to be fanning out on the region that fitted values are greater than 3.0, the general spread looks reasonably constant over the fitted values. The Q-Q plot shows the non-linear trend, but as we have 506 large enough samples, we rely on Central Limit Theorem to satisfy the normality assumption. As we satisfied all 4 assumptions by log transformation through the rest of model selection process, we used log-transformed data.

Results

This is the formula of the final model.

$$\begin{aligned}\log(\text{MEDV}) = & 4.0837 - 0.0103 \times CRIM + 0.0011 \times ZN \\ & + 0.0907 \times RM - 0.0517 \times DIS + 0.0134 \times RAD \\ & + 0.1051 \times CHAS - 0.7217 \times NOX - 0.0006 \times TAX \\ & - 0.0374 \times PTRATIO + 0.0004 \times B - 0.0286 \times LSTAT + \epsilon\end{aligned}$$

Log-linear model

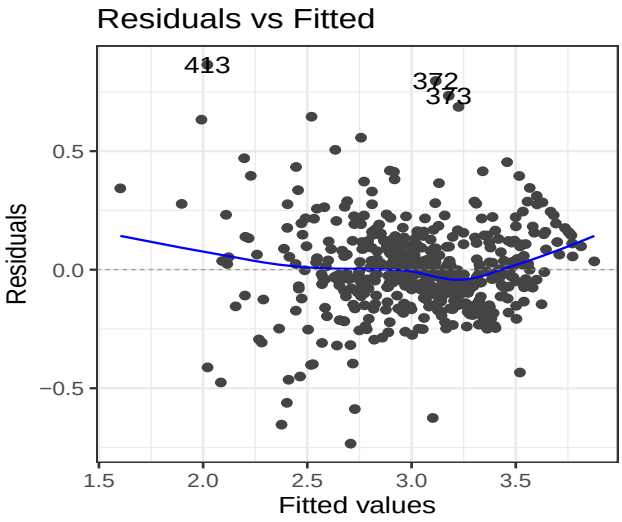
$$log(Y) = \beta_0 + \beta_1 x$$

On average, a one unit increase in x will result in a β_1 * 100% change in Y.

Based on the formula, there are **11 significant elements can effect the price of house**.

- A one degree parts per million(ppm) increase in **NOX** results in **72.17% decrease** in MEDV on average, holding other variables are constant.
- A 1 **CHAS** (if tract bounds Charles River) results in **10.51% increase** in MEDV on average, holding other variables are constant.
- A one number of room increase in RM results in 9.07% increase in MEDV on average, holding other variables are constant.
- A one unit weighted distances to five Boston employment centres increase in DIS results in 5.17% decrease in MEDV on average, holding other variables are constant.

- A one percent increase in PTRATIO results in 3.75% decrease in MEDV on average, holding other variables are constant.
- A one percent increase in LSTAT results in 2.86% decrease in MEDV on average, holding other variables are constant.
- A one index increase in RAD results in 1.34% increase in MEDV on average, holding other variables are constant.
- A one unit of rate increase in CRIM results in 1.03% decrease in MEDV on average, holding other variables are constant.
- A one unit of proportion increase in ZN results in 0.11% increase in MEDV on average, holding other variables are constant.
- A one increase in TAX results in 0.06% decrease in MEDV on average, holding other variables are constant.
- A one 1000(Bk - 0.63)ˆ2 where Bk is the proportion of blacks by town increase in B results in 0.04% increase in MEDV on average, holding other variables are constant.



Discussion and Conclusion

Key conclusions. The increase of criminal rate (CRIM), nitric oxides concentration (NOX), distance to five Boston employment centers (DIS), full-value property-tax rate (TAX), pupil-teacher ratio (PTRATIO), and percentage of lower status of the population (LSTAT) can **decrease** the Boston house price. `NOX` is the most obvious one from the formula.

The increase of proportion of residential land zone (ZN), whether the house was track bounds to Charles River (CHAS), room numbers (RM), accessibility to highways (RAD), racial proportion (B) can **increase** the Boston house price. `CHAS` is the most obvious one from the formula.

As NOX and CHAS is a variable that affects the house price the most, we can infer that the demand for the clean air and nature view was high at the moment that this data was collected, which also aligns with the fact that the data was originally published in the article named "Hedonic prices and the demand for clean air". Also, the increase in room number (represented by RM), decrease of distance to five Boston employment centres(represented by DIS), decrease pupil-teacher ratio by town (represented by PTRATIO) is also a factor that increases the Boston house price.

Limitations. This analysis has 3 limitations. First one is about Assumption for Homoscedasticity. Even though the residuals vs fitted plot showed a fairly random plot, since it had some part mildly fanning out (after 3.5), further test is needed for Homoscedasticity.

Second one is about Bounded MEDV, which is coming from the data itself. The house price (MEDV) was censored at 50.00, that is 50000 dollar. That is, 16 cases that has 50.00 at their MEDV might have a higher price than $50000. Therefore, further research with precise house price is needed to get the model that is better representing the Boston house price.

Last one is related to a small sized data set. The data set has 506 cases, which is large enough to make the model statistically reliable and representative. However, as the dataset sheet argued, it is not enough to represent the whole Boston house price. Increased number of data with better representation on overall Boston house price is needed to make the model more representative.

Although we use 1978 data that is quite outdated, there are some limitations because the current high housing prices hurt many people. We hope that our analysis will help find a balance between people wealth and many factors that can affect house price.

References

GitHub repository

[1]Alan Crawford. (2021, September 20). *The Global Housing Market Is Broken, and It's Dividing Entire Countries*. Available at: https://www.bloomberg.com/news/features/2021-09-19/global-housing-markets-are-hurting-and-it-s-getting-political

[2]Harrison, D. & Rubinfeld, D.L. (1978). *The Boston Housing Dataset*. Available at: https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

[3]Hao Zhu. (2021, February 19). *Create Awesome HTML Table with knitr::kable and kableExtra*. Available at: https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html

[4]R-project.org. (n.d.). *R: The R Project for Statistical Computing*. Available at: https://www.r-project.org/

[5]Rmarkdown.rstudio.com. (n.d.). *R Markdown*. Available at: https://rmarkdown.rstudio.com/

[6]Tidyverse.org. (n.d.). *Tidyverse*. Available at: https://www.tidyverse.org/

[7]Tarr, G (2021). *DATA2002 Data Analytics: Learning from Data*. University of Sydney, Sydney Australia. Available at: https://pages.github.sydney.edu.au/DATA2002/2021/