Housing Price Predicting

Clayton-George Reid

2023-12-07

Introduction

Modeling the relationship between price and other recorded variable will allow for future home buyers to have an understanding of the so called "housing bubble" they are interested in. The analysis and model used in this paper suggests that the price is related to the number of bedrooms, number of bathrooms, square foot living space, square foot living of near by houses, grade(type of architecture and level of construction) ,and if it has a waterfront view. Individual who plan on buying a home in the future can use this model in this paper to predict an estimated price the home they are interested in will go for sale.

Methods

Data Collection

Data used in this paper was originally downloaded from https://lasanthi-asu.github.io/STT3851ClassRepo/Rmarkdown/Data/housedata.csv. As for when or whom created this data set, I would have to assume it was either my Professor Watagoda or she plucked it from the internet. Research to find the orginal data set was not successful as there are multiple other data sets called house data.

Statistical Modeling

Standard multivariate regression techniques such as ordinary least squares (OLS) regression, linear regression,and logarithmic transformation.

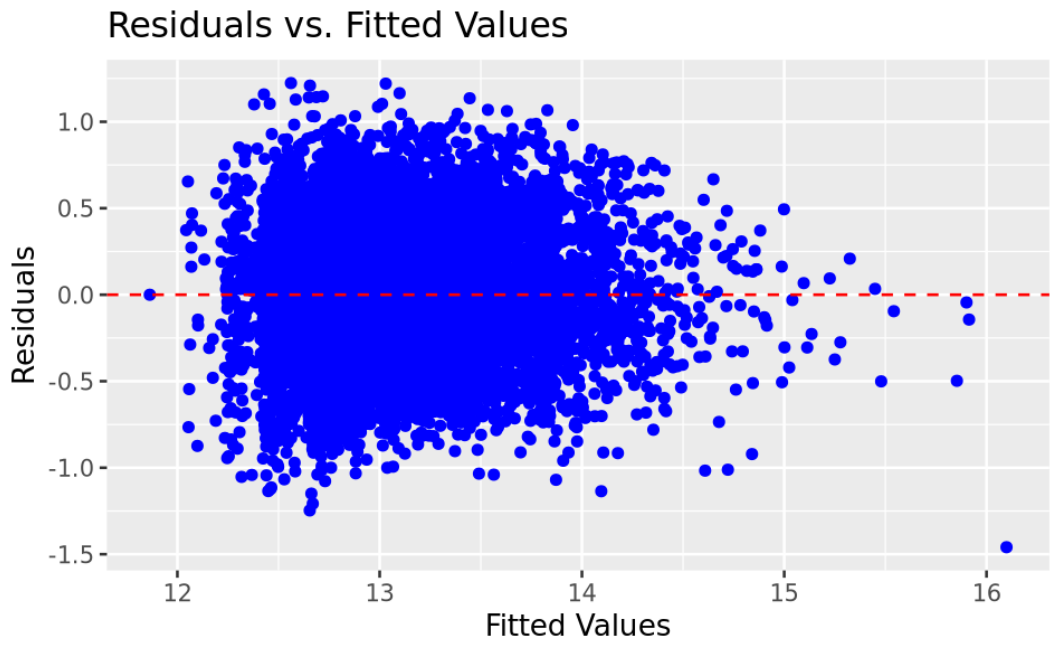
Reproducibility

All analyses performed in this paper can be reproduced by running the original .Rmd file with RStudio, assuming the link to the original data remains current and the contents thereof remain unchanged. The R packages ggplot2 (Wickham et al. 2020), knitr (Xie 2021),and rmarkdown (Allaire et al. 2021),will need to be installed on the user's computer.

Results

The data used to create the final model included number of bedrooms, number of bathrooms, square foot living space, square foot living of near by houses, grade(type of architecture and level of construction) ,and if it has a waterfront view. Quick note the price variabele when creating the model was log(price). The model used was:

$$lm(log(price)\ bedrooms + bathrooms + sqft_living + sqft_living15 + grade + waterfront, data = housedata)$$



Title: "Residuals vs. Fitted Values" X-axis: "Fitted Values" - Represents the predicted or fitted values. Y-axis: "Residuals" - Represents the differences between the observed and predicted values.

Graph Features:

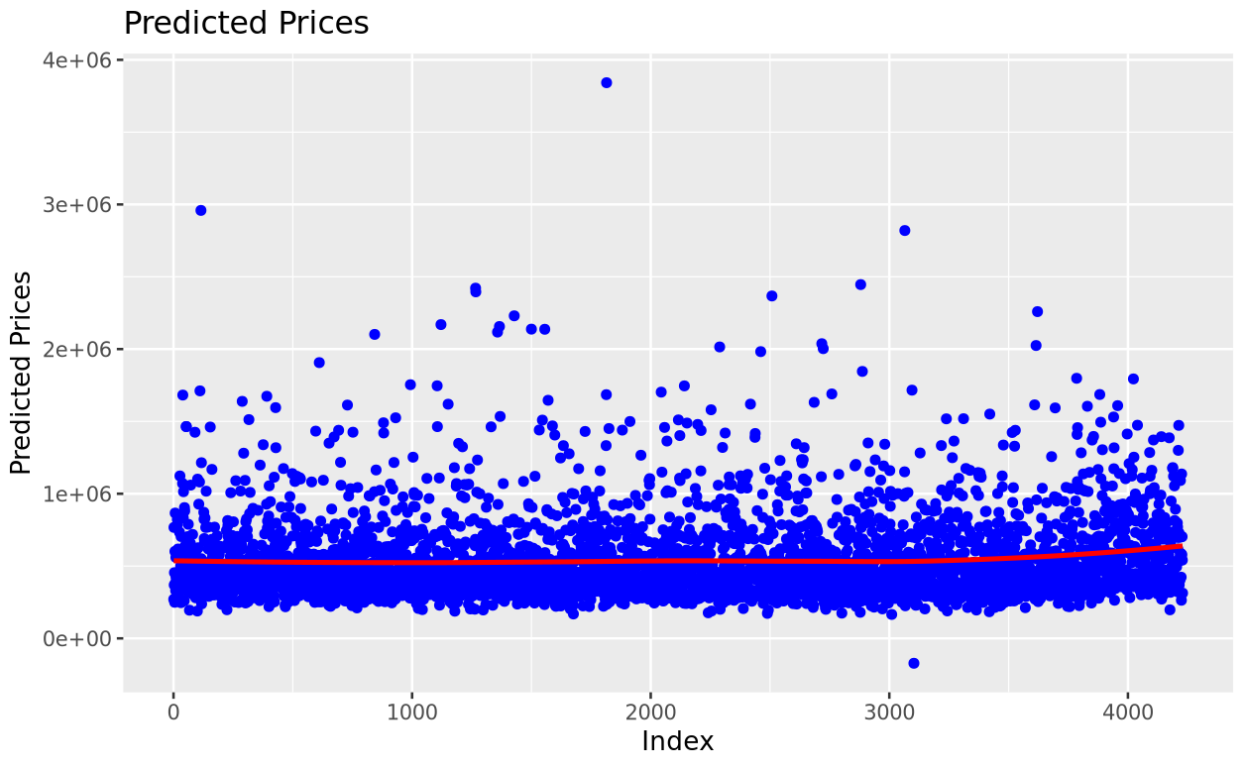
Scatter Points (Blue): Each blue point on the graph represents a data point. The x-coordinate is determined by the fitted values, and the y-coordinate is determined by the corresponding residual value from the model.

Dashed Red Line at y = 0: Represents the line where residuals are equal to zero. It serves as a reference to assess whether the model is making unbiased predictions.

Interpretation:

The majority of points concentrated around the y = 0 line suggest that, on average, the model is making unbiased predictions. However, the persistence of a funnel-shaped pattern may indicate that the variability of residuals is not constant across different levels of fitted values.

If the points are still densely packed and do not exhibit a clear pattern of spread around the zero line, it could suggest that there may be inherent variability in the data that is not effectively captured by the current model.



Title: "Predicted Prices" X-axis: "Index" - Represents the index variable. Y-axis: "Predicted Prices" - Represents the predicted prices from the Reid_ClaytonGeorge variable in the housedataT dataset.

Graph Features:

Scatter Points (Blue): Each blue point on the graph represents a data point from the housedataT dataset. The x-coordinate is determined by the "Index," and the y-coordinate is determined by the corresponding value in the "Reid_ClaytonGeorge" variable.

Smoothed Line (Red): The red smoothed line is generated using the loess method. It represents the trend in the data, providing a smoothed curve that captures the general pattern in the relationship between the "Index" and "Reid_ClaytonGeorge" variables. The line is not a straight line but flexes to better fit the data.

Interpretation:

- The densely packed blue points around the intercept indicate that a majority of the data points cluster closely around a certain predicted price value at the intercept.
- The red smoothed line provides a visual representation of the overall trend in the data. If the line is curving upwards or downwards, it suggests a non-linear relationship between the "Index" and "Reid_ClaytonGeorge" variables.

Conclusions

The goal of this analysis was to produce a model that can predict the price of a future home, in this analysis, a linear regression model was developed to explore the relationship between various housing features and property prices. The initial model revealed a heteroscedastic pattern in the residuals vs. fitted values plot, indicating that the variance of errors was not constant across all levels of predicted values.

To address this issue, several strategies were employed, including a logarithmic transformation of the response variable and the exploration of weighted least squares and robust regression techniques. Despite these

attempts, the residuals vs. fitted values plot still exhibited a concentrated pattern around the zero line, suggesting persistent challenges with heteroscedasticity.

Further refinements, such as investigating influential points, exploring additional predictor variables, or considering alternative modeling approaches, may be necessary to enhance the model's fit. It's crucial to carefully evaluate model assumptions and diagnostic plots to iteratively improve the regression model and ensure reliable predictions.

```
This document uses DT by @R-DT, ggplot2 by @R-ggplot2 by , plotly by @R-plotly, rmarkdown by @R-rmarkdo
```

```
The previous line with citations was created using:
```

This document uses `DT` by Xie, Cheng, and Tan (2023), `ggplot2` by Wickham et al. (2022), `ISLR` by James et al. (2021), `plotly` by Sievert et al. (2023), `rmarkdown` by Allaire et al. (2023), `dplyr` by Wickham et al. (2023), `knitr` by Xie (2023b), and `bookdown` by Xie (2023a). "

```
sessionInfo()
```

```
R version 4.2.3 (2023-03-15)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Red Hat Enterprise Linux 9.2 (Plow)

Matrix products: default
BLAS/LAPACK: /usr/lib64/libopenblasp-r0.3.21.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] knitr_1.45

loaded via a namespace (and not attached):
 [1] compiler_4.2.2    fastmap_1.1.1     cli_3.6.1         tools_4.2.2
 [5] htmltools_0.5.7   rstudioapi_0.15.0 yaml_2.3.7        rmarkdown_2.25
 [9] xfun_0.41         digest_0.6.33     rlang_1.1.2       evaluate_0.23
```

References

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2023. *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

James, Gareth, Daniela Witten, Trevor Hastie, and Rob Tibshirani. 2021. *ISLR: Data for an Introduction to Statistical Learning with Applications in r*. https://www.statlearning.com.

Sievert, Carson, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, and Pedro Despouy. 2023. *Plotly: Create Interactive Web Graphics via Plotly.js*. https://plotly-r.com.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2022. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Xie, Yihui. 2023a. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://github.com/rstudio/bookdown.

———. 2023b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Xie, Yihui, Joe Cheng, and Xianying Tan. 2023. *DT: A Wrapper of the JavaScript Library DataTables*. https://github.com/rstudio/DT.