Modelling New York House Prices

Josh Bercich[1], **Ryan Dharma**[1], **Raymond Huang**[1], **Hayden Savage**[1], **and Ben Shade**[1]

[1]University of Sydney

This version was compiled on November 20, 2020

This report examines a number of predictive multiple regression models in an attempt to derive a robust yet usable model for homeowners and other interest groups. Models were created with OLS regressions on house prices using a 2006 New York house price dataset. In an attempt to find a better model, log-linear models were also created. It was concluded that a seven parameter linear model produced comparatively accurate predictions, consisting of land value, living size, lot size, number of bathrooms, and whether the property is a waterfront, new construct or has central heating. Despite this, a number of limitations restricts its applicability in the modern day.

Introduction

Houses are often the single largest component of an individual's personal assets. They impact consumer behaviour, household debt and the distribution of wealth. We thus sought to create a model for valuing houses based on a number of explanatory variables that were both statistically significant and accessible to individuals. In this way, it can be used by developers, appraisers, and potential homeowners to estimate prices, assist in negotiations, and identify overpriced or undervalued properties. Furthermore, this model could assist in the maintenance of property price indices, which have use cases for policy makers.

Data set

Our data set contains a random sample of 1734 houses from Saratoga County, New York in 2006. This sample comes from the Saratoga Housing Data (De Veaux). No details were given about the methodology behind the data collection.

The data itself contains house prices in US dollars as well as 16 other quantitative, categorical and dummy variables. A full description of the variables can be found in the appendix.

To clean the data, dummy variables were initially switched to categorical "Yes" and "No" variables so that the `lm()` function did not treat them as numeric. The only observations that appeared to be incorrect were those with `Lot.Size` equal to 0, and these were removed. We also binned the categories for `Fuel.Type` that weren't `Oil`, `Electric` or `Gas` into an `Other` category, as the original categories contained very few observations. Furthermore, the variable `Pct.College` was removed since the percentage that graduated college in a neighbourhood would be very difficult for ordinary users to determine. Finally, the variable `Test` was removed due to a lack of metadata describing it.

It should be noted that some observations may not be independent with each other since houses in a specific geographic area share similar structural features. This issue has been taken into consideration, although random sampling should generally alleviate its impacts.

Analysis

Two models were made: an OLS regression of Price against a number of regressors, and a regression of log(Price) against a number of regressors.

Linear model (no transformations)

Initially, linearity was checked to avoid any systematic shortcomings in the model. Pairwise plots of price against each variable showed a clear relationship for the living area, land value, number of rooms and bathrooms. As for the number of fireplaces and bedrooms, their deviation can be attributed to the small number of discrete values, and thus they were retained. Age and lot size were contentious for this assumption but as commonly accepted regressors of property prices, they were also retained.

We began by regressing price against all variables to create a full model. We then compared this to the models generated by two algorithmic approaches: backwards selection using p-values, and backwards selection using AIC. Both approaches generated the same model, which depended on 11 variables.

Both the full model and the backwards-step AIC model met the necessary assumptions to a reasonable extent. The residual plots for both models showed the residuals were fairly linear and homoskedastic. The points in the Q-Q plot deviate from the normal quantiles at the extremes, indicating that the residuals may not be normally distributed. However, due to the large sample size, the Central Limit Theorem can be relied upon to ensure the inferences are approximately valid. The aforementioned residual plots can be found in the appendix.

These initial models regressed on a large number of variables, which reduces their accessibilty to ordinary users and runs the risk of overfitting. Thus, we sought to find other, smaller models by examining model stability and variable selection plots.

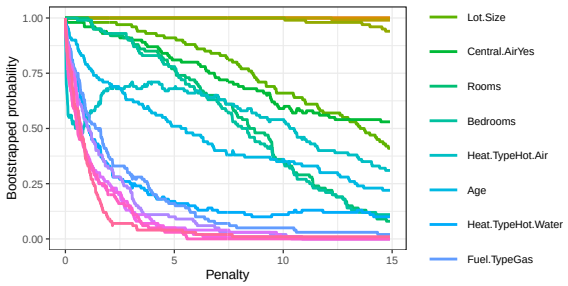

Fig. 1. Variable inclusion plot for regression penalties with Price using the full model space.

The variable selection plot shows that living area, land value, waterfront and bathrooms are highly important estimators, and that lot size and central air may also be worth considering. We then examined the model stability plot, which is shown below:

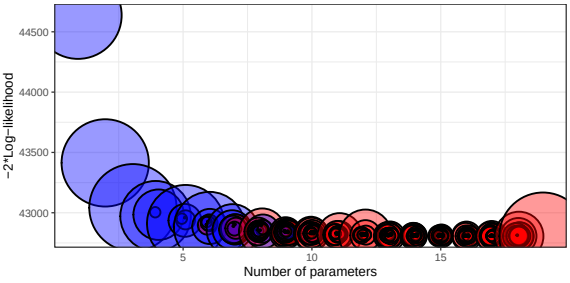

Fig. 2. Bootstrapped model stability plot of the full model space representing model probabilities.

Analysis of this plot led us to propose two new models:

- A 4-parameter model consisting of 'Waterfront', 'Land.Value', 'Living.Area' and 'Bathrooms'.
- A 7-parameter model consisting of 'Waterfront', 'Land.Value', 'New.Construct', 'Living.Area', 'Bathrooms', 'Central.Air' and 'Lot.Size'.

A comparison of the performance of our candidate models is shown in the table below:

Model	RMSE	MAE	AdjRsquared
Full model	56879	41489	0.658
Backwards AIC	56625	41464	0.663
7-Parameter	57160	41858	0.662
4-Parameter	58190	42548	0.643

From this table, it is clear that the backwards AIC model is the best-performing, as it has the lowest RMSE and MAE, and the highest adjusted R-squared. However, the difference is only marginal. Furthermore, the plots below show that for the AIC and 7-parameter models, there is substantial overlap in the confidence intervals for the cross validation error rates.

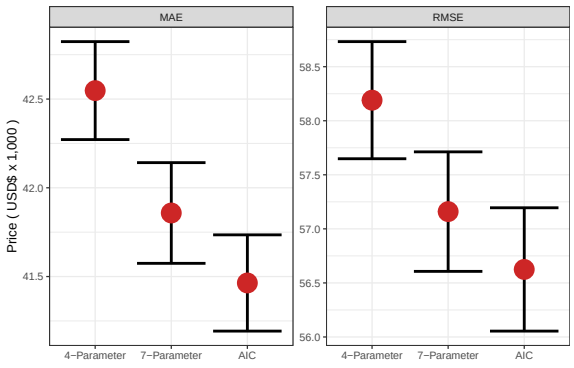

Fig. 3. Confidence intervals for each error measure between 4-parameter, 7-parameter and AIC models.

Due to this overlap, selecting the 7-parameter model over the AIC can be justified, as the marginally poorer performance is outweighed by the benefits of having a smaller, more accessible model. Thus, our chosen model is given by the equation:

$$\begin{aligned}\text{Price} = &-311.61 + 121,207.19(\text{Waterfront}_{\text{Yes}}) \\ &+ 0.91(\text{Land.Value}) - 37,639.53(\text{New.Construct}_{\text{Yes}}) \\ &+ 70.05(\text{Living.Area}) + 25,798.37(\text{Bathrooms}) \\ &+ 15,995.61(\text{Central.Air}_{\text{Yes}}) + 7,386.96(\text{Lot.Size}) + \epsilon\end{aligned}$$

Assumption checking for this model is in the appendix.

Log-linear model

In an attempt to improve our model, we tried regressing on the log of the house prices.

A similar approach to above was used for model selection, whereby a full model, a backwards-step AIC model, and a backwards-step p-value model were created. The variable selection and model stability plots were then examined, and upon performance comparison, we settled upon a 4-parameter model as our chosen log-linear model.

The log-linear model appeared to preserve the linearity of the pairwise relationships. However, the homoskedasticity and linearity of the residuals were negatively affected, particularly for large house price values. The deviations from the Q-Q line at the extremes were similar to those in our original linear model.

Our chosen log-linear model had a higher RMSE and MAE than our original model, as well as a lower adjusted R-squared value. Thus, it was concluded that the log transformation of price did not improve the quality of our model. Hence, the 7-parameter model described earlier is our final choice of regression model.

Results

As evident in the equation for our model, all but two of the regression coefficients are positive. In particular, we can see that a waterfront property will cause the house price to increase $121,207, and central air conditioning increases house price by $15,996. A $1 increase in land value will result in a 91c increase in price, while an extra bathroom increases house price by $25,798. Furthermore, 1 square foot increase in living area will result in a $70 increase in price, and a 1 acre increase in lot size will result in a $7,387 increase in price.

Interestingly, labeling a house as a new construct resulted in a $37,640 reduction in price. This is an unusual result, as one would expect that holding all else constant, houses that were more recently constructed should have higher prices. To determine if there were any variable interactions causing the sign to be negative, we used a two-factor ANOVA. When using a Bonferroni correction, the only significant interaction we found was with `Central.Air`, however this interaction was not attested to by our variable inclusion plot. Additionally, through data exploration we found there were no new constructions that were also waterfronts, which may be causing an issue given the large positive coefficient for `Waterfront`. Use of a different and preferably larger data set would allow us to gain a better understanding.

The R-squared value of the model shows that the seven variables used explain approximately 64% of the variance in price.

Discussion and conclusion

One limitation was that a large amount of variables were spatially autocorrelated due to neighbourhoods having similar structural characteristics - this can violate the assumptions of independent observations and homoskedasticity (Basu & Thibodeau, 1998). This could be improved in future research by incorporating the observed spatial relationships. Furthermore, multicollinearity between coefficients related to each other can reduce the statistical power of the model. This is because it becomes harder to determine which variables are statistically significant, making it more difficult to justify which is the best model. Investigating these relationships in the future can allow for better model selection and justification.

Furthermore, the dataset is quite old as it is from 2006. If we are trying to use this model to predict current house prices, we would need to adjust for temporal effects like inflation and exogenous macroeconomic shocks that could have large implications for house prices. Finally, there are limited applications for this model to Sydney, given that these are New York house prices. Investigating other problem domains may help make the model more general (Park & Bae, 2015).

Nonetheless, by investigating numerous models using OLS regression, we were able to find a model that can reasonably predict house prices using seven predictor variables. From this, we found that the features of a house that had the greatest impact on its price were land value, living size, lot size, number of bathrooms, and whether the property is a waterfront, new construct or has central heating.

Appendix

Variable descriptions:

Quantitative Variables: `Lot.Size` (acres), `Age` (years), `Land.Value` (USD), `Living.Area` (sq. ft), `Pct.College` (percent of neighbourhood that graduated college), and the number of: `Bedrooms`, `Fireplaces`, `Bathrooms`, `Rooms`.

Dummy Variables: whether the property is a `Waterfront`, `New.Construct` or has `Central.Air`.

Categorical Variables: `Fuel.Type`, `Heat.Type`, `Sewer.Type`.

Linearity plots: Linearity of `Age` and `Lot.Size` against `Price`. Fanning of points about the regression line contradicts linearity but there exists an implied relationship between age and price. Similarly, lot size and price are expected to share a positive correlation but due to the wide range of prices for the densely populated lot sizes, this also compromises linearity, as supported by the growing confidence interval.

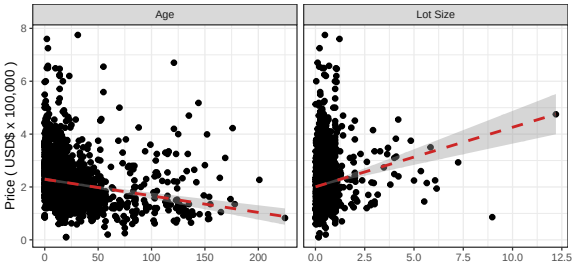

Fig. 4. Linearity of property age and lot size against Price.

Final 7-parameter model residual and Q-Q plot: Linearity and homoskedasticity assumptions appear to be met, with a fairly uniform distribution of residuals above and below the x-axis, as well as an approximately equal spread of residuals within the plot. Normality of the Q-Q plot is contentious, but due to the sample size and CLT, our inferences are approximately valid.

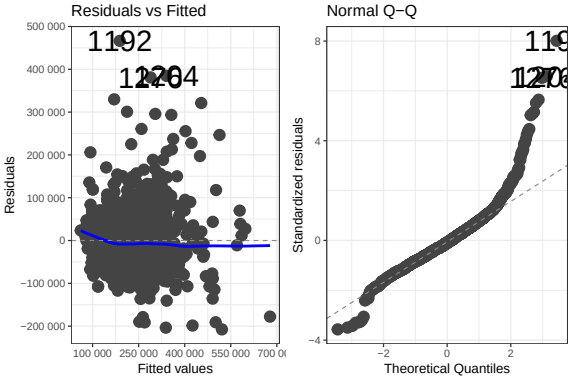

Fig. 5. Final 7-parameter model residual plot and Q-Q plot.

Full model residual and Q-Q plot:

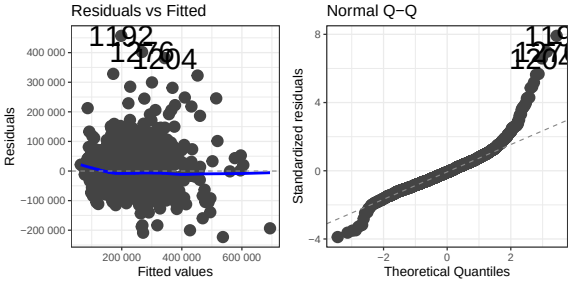

Fig. 6. Full model residual plot and Q-Q plot.

Backwards AIC model residual and Q-Q plot:

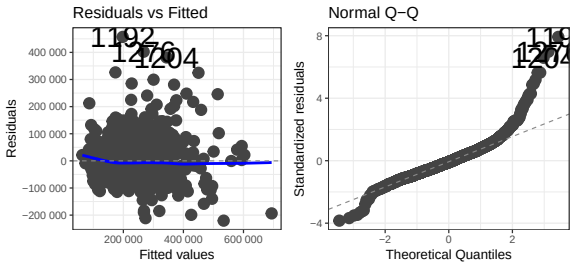

Fig. 7. Backwards AIC model residual plot and Q-Q plot.

4-parameter log-linear model residual and Q-Q plot:

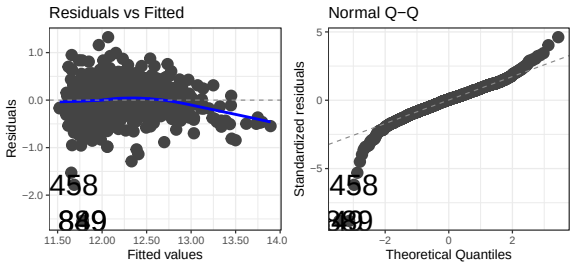

Fig. 8. Log-linear model residual plot and Q-Q plot.

References

- Basu, S., Thibodeau, T., G. (1998) Analysis of Spatial Autocorelation in House Prices. Journal of Real Estate Finance and Economics, 17(1), 61-85. doi: https://doi.org/10.1023/A:1007703229507
- Broman KW (2015) R/qtlcharts: interactive graphics for quantitative trait locus mapping. Genetics 199:359-361 doi:10.1534/genetics.114.172742
- Daniel Anderson and Andrew Heiss (2020). equatiomatic: Transform Models into 'LaTeX' Equations. R package version 0.1.0. https://CRAN.R-project.org/package=equatiomatic
- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom
- Hadley Wickham (2020). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.0. https://CRAN.R-project.org/package=forcats
- Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
- Masaaki Horikoshi and Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify
- Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934. doi: https://doi.org/10.1016/j.eswa.2014.11.040
- Thomas Lumley based on Fortran code by Alan Miller (2020). leaps: Regression Subset Selection. R package version 3.1. https://CRAN.R-project.org/package=leaps
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Git: https://github.sydney.edu.au/bsha5224/T09oc_early_2