Linear Regression on New York Housing Data

Sourish Iyengar

Multiple linear regression was used to attempt to accurately, yet robustly predict house prices with given attributes. Our model found that the presence of a waterfront, central air-conditioning, construct type and the number of bathrooms was associated with a large impact on house price. Unexpectedly, new constructs have a negative influence on price. This requires further examination. Our final model of 9 predictors out of 16 had a MAE of 41443, the best out of our tested models. Limitations and additional variables that may improve performance are identified.

Introduction & Dataset Description.

When buying, developing or investing in properties, it can be difficult to decide what factors are important in maximising affordability or returns. Furthermore, house appraisals can be time consuming, expensive and prone to human error. Thus, our analysis aims to determine a systematic relationship between variables found in a housing dataset and their price using a multiple linear regression model.

The dataset used was collected by Candice Corvetti, and is a random sample of 1734 houses in Saratoga County, New York in 2006. It contains 16 explanatory variables that relate to the size, age, room types, land value and categorical features of a house. These variables and their abbreviated names are described in A1.

Initial Filtering and Variable Transformations.

House prices by category was observed. From A2, not having heating or fuel type is associated with lower prices relative to other categories for their respective variables. To aid with model interpretability, they were collapsed into binary variables that do or do not have heating/fueling. From A3,the prices across categories in the test and sewer type variables didn't seem to differ drastically. Thus, these variables are not fitted in the following models. The other categorical variables are visualised in A4.

Relative to the age, $log(age + 1)$ had a more pairwise linear relationship with price and thus was transformed. From A5, lot size, bedrooms and percentage college were dropped as they did not seem to be linear with price even after using a log transformation.

All the assumptions of linear regression were satisfied on the full model. These will be verified again after the final model is selected.

Model Selection.

Stepwise Regression.

Backward and forward stepwise regression was performed. This involves a non-exhaustive construction of a regression model that iteratively drops (backward selection) or adds (foward selection) parameters to minimise the AIC score. Both methods of stepwise regression dropped fuel type and fireplace from the final model.

Model Stability.

The variable inclusion plot varies the value of the penalty coefficient, λ, on the x-axis. On the y-axis, the probability a parameter is included in models that minimise the loss plus penalty in 150 weighted bootstrap samples is observed. This is done to see the effect of slight deviations in our samples in the selection of various parameters. If stable, the probability a parameter is selected should be relatively consistent as λ is varied.

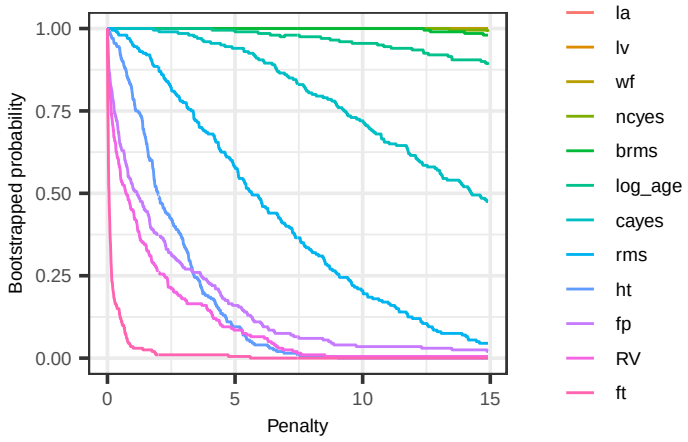


Figure 1: Variable Inclusion Plot

In Figure 1, `RV` is a redundant variable independent of price. Lines following a similar path to `RV` such as `ft`,`ht` and `fp` aren't likely to be in a stable model. Conversely, `la`, `lv`, `wf`, `nc`, `brms`, `ca` and `log_age` are selected with relatively high probability throughout. The probability that the `rms's` variable is selected converges to 0 as λ increases. However, this occurs relatively slowly. This is investigated using model stability plots (MSPs).

In MSPs, $\lambda = 2$ is fixed, a circle represents a model and its size is proportional to its stability in the bootstrap samples.

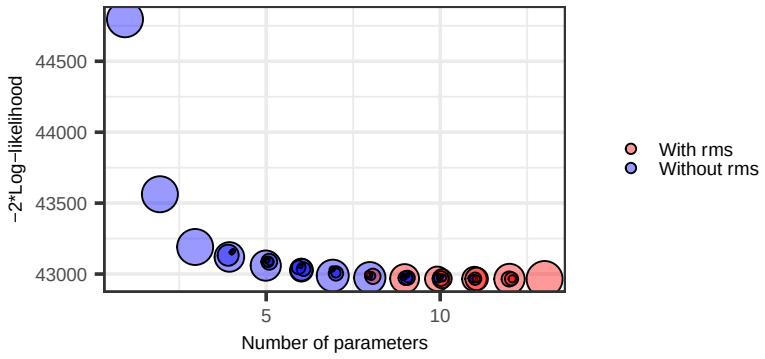


Figure 2: Model Stability Plot

From Figure 2, the model selected through stepwise regression isn't overwhelmingly dominant in its dimension, $k = 10$. Contrastingly, rooms is selected in a clear dominant model of $k = 9$. There are also stable models of $k = 5,7,8$ with comparable probabilities of selection. However, this is empirically stronger for the model of size 8 as it is in a higher dimension.

We proceed with the stable dominant models of $k \in 8,9$ for model evaluation.

Model Evaluation - K-fold Cross Validation.

Model	R^2	RMSE	MAE
Stepwise	0.6561857	58422.57	41458.40
Stable (K=8)	0.6554816	58351.42	41609.18
Stable (K= 9)	0.6496195	58122.72	41442.76

Figure 3: 10-fold CV-Out of Sample Performance

Figure 3 shows that on average, the dimension 9 stable model had the best out of sample performance in 2/3 metrics. Although it had the lowest R^2 value, relative to the stepwise and dimension 8 model, it performs better regardless of whether large errors are penalised to a greater extent. This is indicated by its MAE and RMSE score. Compared to R^2, RMSE and MAE are easier to interpret for non-statisticians. Thus, we chose to proceed with the dimension 9 model as it performs the best for these metrics.

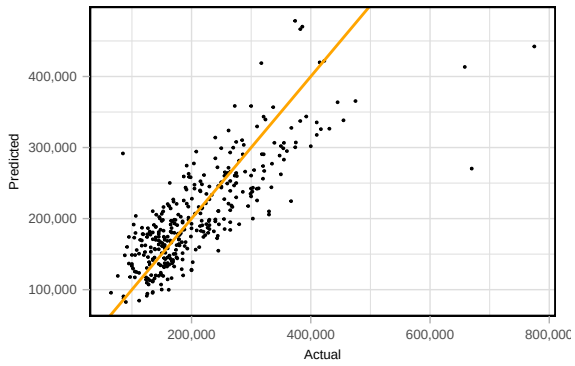


Figure 4: Out of sample Actual vs Predicted Plot

Figure 4 displays the out of sample performance of the stable dimension 9 model on 20% of the data. The position of the three houses on the far right indicate that they may be overpriced, relative to the other houses. Further exploration of these observations may reveal other factors that influence house prices.

Assumptions of the Final Selected Model.

Assumptions of linear regression must be satisfied to ensure it is appropriate to fit a linear model and inferences are valid. They are linearity between predictors and price, homoscedasticity, independence and normality in the error distribution.

From A6, there is pairwise linearity between predictors against price. In A7, the residuals are symmetrically distributed about 0, indicating multivariate linearity is satisfied.

The residuals in A7 with a value above 250,000 may indicate heteroskedasticity. These points are a small proportion of the dataset and are unlikely to have a detrimental impact on inferences. Thus, we proceed with this assumption being satisfied.

In A7, the residuals don't fall on the lower and upper tail of the qqnorm diagonal line, indicating non-normality. However, since the sample size is large ($n = 1734$), we rely on the Central Limit Theorem for approximately valid inferences.

Independence is assumed by the method of data collection.

Interpreting Coefficients.

$$\begin{aligned}\text{price} = 33317.24 + 67.21(\text{livingarea}) + 19367.65(\text{bathrooms}) \\ +0.93(\text{landvalue}) - 9226.81(\log(\text{age})) + 122570.59(\text{waterfront}_{\text{yes}}) \\ -59575.95(\text{newconstruct}_{\text{yes}}) + 12562.28(\text{centralair}_{\text{yes}}) \\ +2183.2(\text{rooms}) + \epsilon\end{aligned}$$

The intercept and all coefficients were found to significantly differ from 0 at a 5% significance level. The intercept of the final model is not interpretable in the context of the domain, as a house cannot exist when all the other parameters are 0.

There are positive coefficients for living area, bathrooms, land value, waterfront, central air conditioning and rooms. In this model, these are the distinguishable features of higher end houses. As expected, an increase in one dollar of land value is approximate to one dollar of price. Comparing the magnitude of coefficients reveals that bathrooms have a stronger weight than rooms with regards to the house price. The majority of homes have more rooms than bathrooms and this is reflected in the model. Additionally, waterfront, new construct, bathrooms and central air conditioning are associated with a relatively high magnitude of impact on price.

There are negative coefficients for log age and new construct. A 1% increase in age is correlated with a $92 decrease in price. Interestingly, new constructs lower the price of the house. This may be confounded by other underlying variables such as house type or location, as newer constructions are typically multi-dwellings or built further from the urban centre.

Limitations, Conclusion and Further Study.

Since the data was sourced in 2006, our analysis may not be valid in 2020. House prices also tend to fluctuate with global economic conditions but this is not accounted for in our model. Additionally, the data only contains samples from houses in Saratoga, so our analysis may not apply to outside this region. It would also be inappropriate to predict the value of houses with attributes differing from what the model was trained on.

In conclusion, we developed a stable final model that that performs fairly well in predicting house prices and may be fine-tuned further prior to deployment. An appropriate further investigation would be the location and type of newly constructed homes. We could scrutinise the abnormal points from Figure 4 to gain further insight. Additionally, having a wider range of attributes such as proximity to public transport,schools, the neighbourhood crime rates, and more housing data beyond Saratoga, would allow for a more robust model to be developed.

References.

Corvetti, C. (2020). SaratogaHouses: Houses in Saratoga County (2006). mosaicData: Project MOSAIC Data Sets. https://rdrr.io/cran/mosaicData/man/SaratogaHouses.html

Kuhn, M. (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

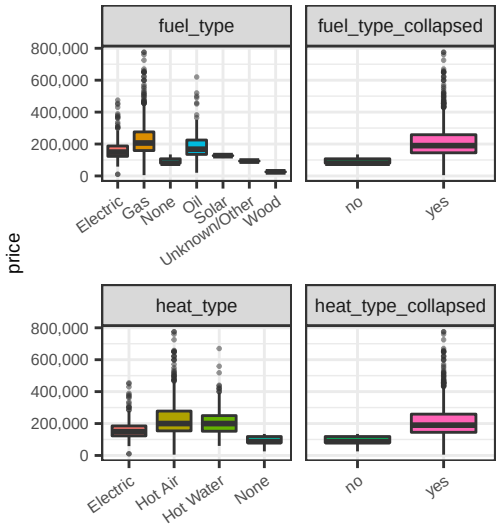
Tarr G, Müller S, Welsh AH (2018). "mplot: An R Package for Graphical Model Stability and Variable Selection Procedures." *Journal of Statistical Software*, *83*(9), 1-28. doi: 10.18637/jss.v083.i09 (URL: https://doi.org/10.18637/jss.v083.i09).

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

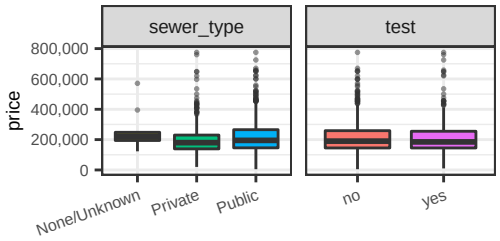
Appendix.

Variable		Description
Numeric	Price	Price of the house (US dollars)
	Lot Size	Size of the lot (acres)
	Age	Age of the house (years)
	Land Value **(lv)**	Value of the land (US dollars)
	Living Area **(la)**	Living area (square feet)
	% College	Percent of neighbourhood that graduated college
	Bedrooms	Number of bedrooms
	Fireplaces **(fp)**	Number of fireplaces
	Bathrooms **(brms)**	Number of bathrooms
	Rooms **(rms)**	Number of rooms
Categorical	Heating Type **(ht)**	Type of heating system
	Fuel Type **(ft)**	Fuel used for heating
	Sewer Type	Type of sewer system
	Waterfront **(wf)**	Whether the property includes a waterfront
	New Construction **(nc)**	Whether the property is a new construction
	Central AC **(ca)**	Whether the property has central air-conditioning
	Test	Dummy variable

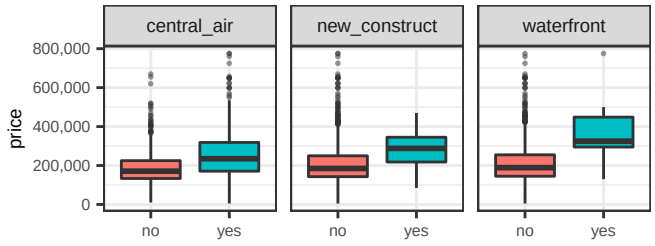
A1. Dataset Variable Descriptions



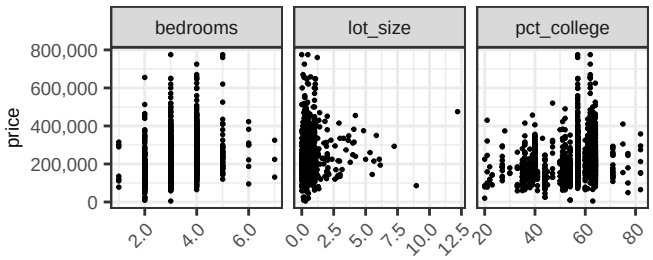
A2. Transformed fuel_type and heat_type



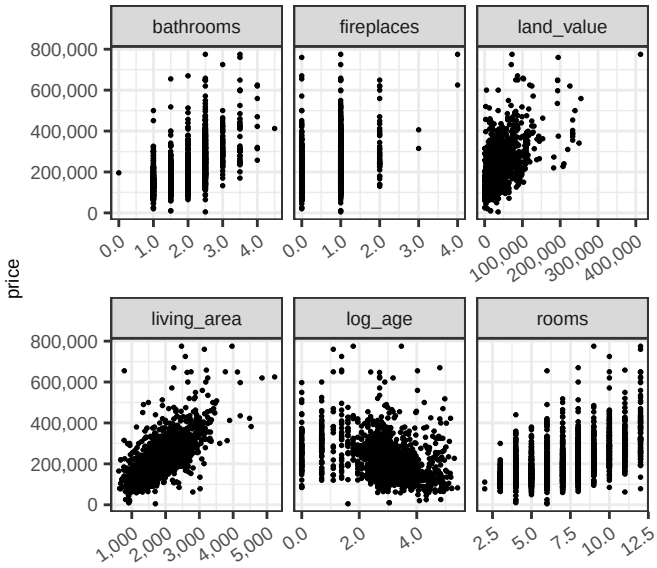
A3. Dropped Categorical Variables against Price



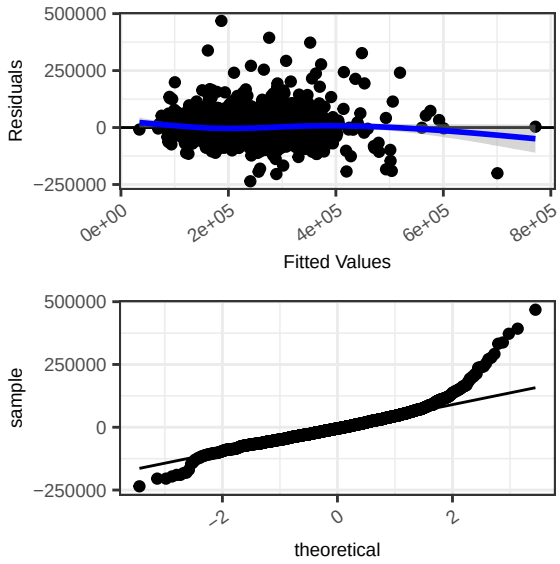
A4. Other Categorical Variables against Price



A5. Dropped Numerical Variables against Price



A6. Final Model Numeric Variables against Price



A7: Residual v.s. Fitted Plot and Normal QQplot of the Residuals