Ames Iowa House Prediction Project

Jordan Kassof and Jose Torres

2/11/2018

Contents

Introduction

Buying a house is one of the largest financial decisions many people will make. So many factors go into someone's decision, but it can be hard to really explain why one house "felt right" and another didn't. We want to quantify the factors that add up to someone making the decision to purchase a house.

Data Description

We will be using the Ames, Iowa individual residential property sales data set freely available on Kaggle.com. The dataset contains 2,930 observations with 79 explanatory variables. All observations occur between 2006 and 2010. Given the geography dependent nature of home value, the results of below analyses can't be applied nationally. For more information on the data, or to download it yourself, visit https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

See the codebook.txt file in the github repo for complete information about all variables.

Exploratory Analysis

Comments on various things noticed as exloring the data and general idea of what the cleaning was. Reference appendix for entire cleaninf functions and scripts.

Questions of Interest

We focused on two approaches, one geared towards model performance and one towards model interpretability by one of the parties involves in a home purchase.

Interpretable Models

For the interpretable model approach, rather than just taking a handful of easily understood parameters and building a model, we wanted to take a different approach. There are many adages when it comes to home buying, we wanted to see which are the most true. The three ideas about what drives the price of a house that we looked at are as follows:

- Location, location, location!
  - For this model, we used parameters that are related to the physical location of the property. For example, neighborhood, zoning, frontage, lot size, etc.
- It's all about the curb appeal
  - For this model, we used parameters related to the external appearance of the property. For exampe, house style, roof style, external vaneer materials, etc.
- It's what's on the inside that counts
  - For this model, we used parameters related to the internals of the property, the bones if you will. For example, the foundation, the electrical and heating system, etc.

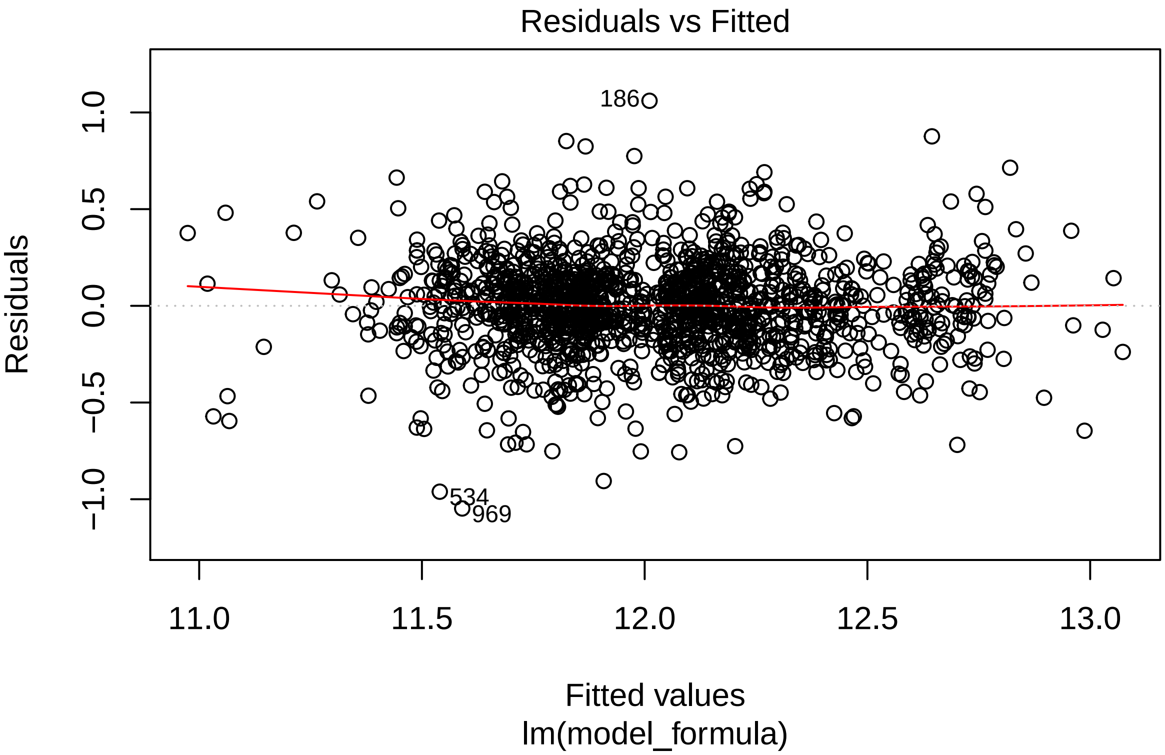
We also included the Sale Condition variable in all three models because the context of the sale seems like way too key of a factor to leave out of any model that is meant to be easily interpretted.

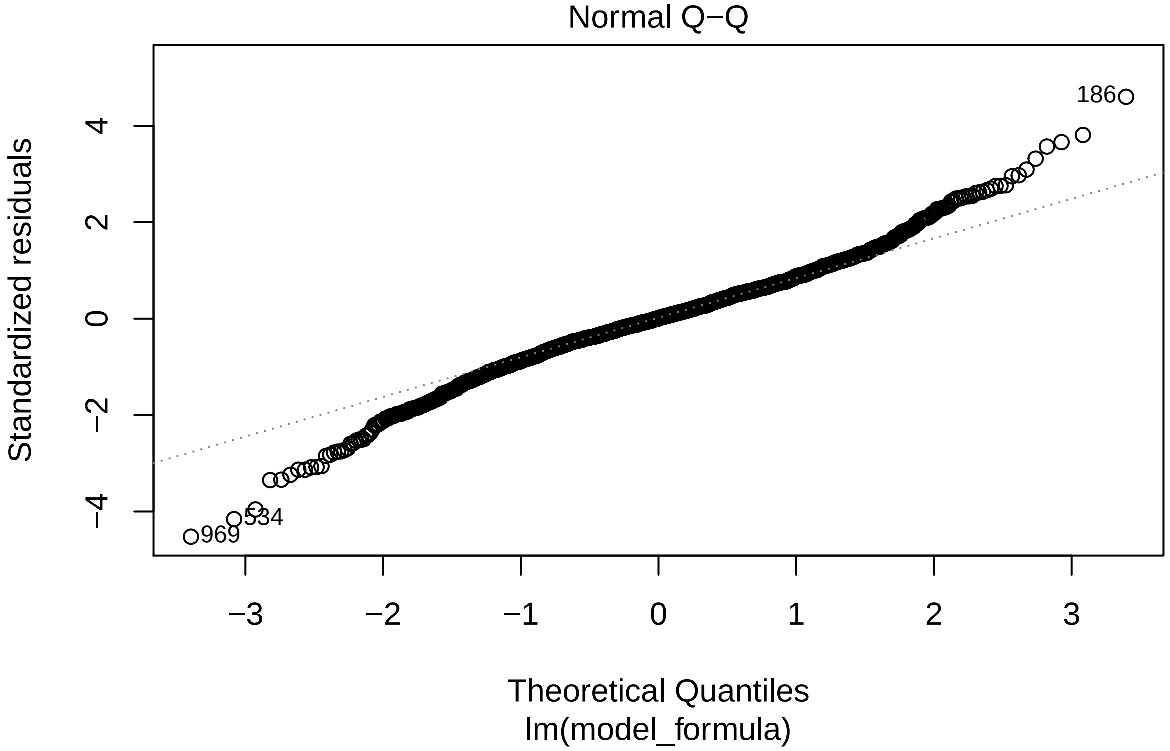
Model Selection

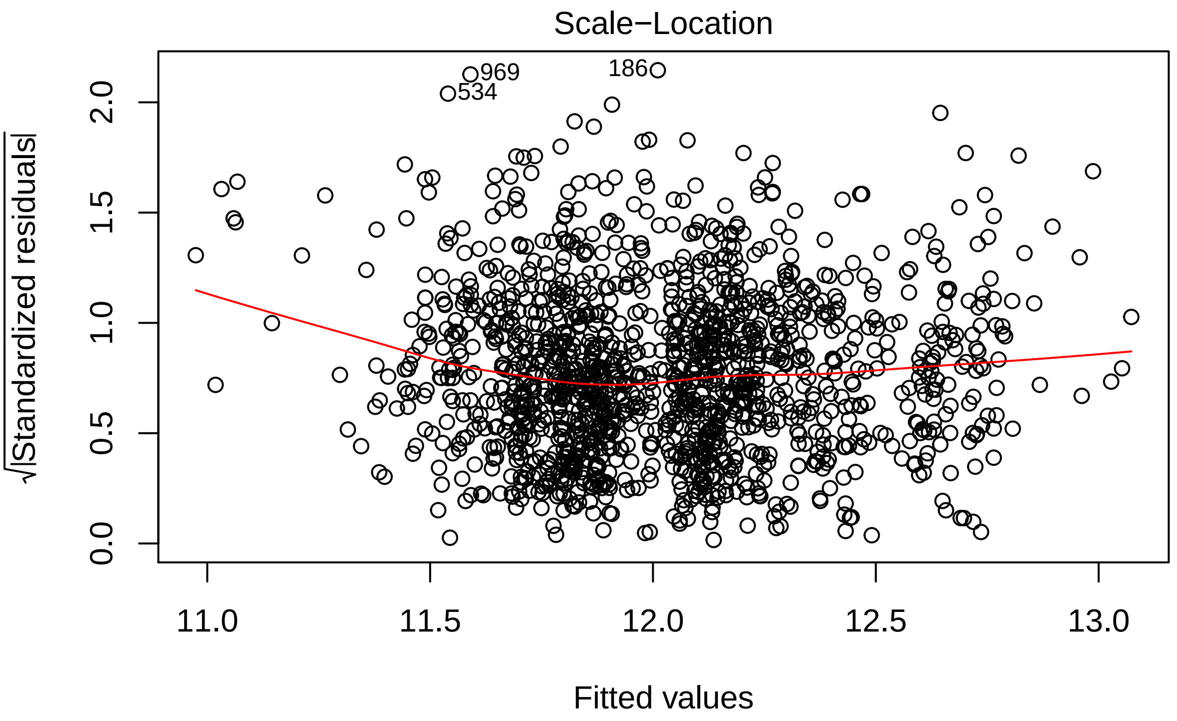
After running our three models, let's take a look at some diagnostics. After picking a model based on diagnostics, we will examine assumptions and parameters.

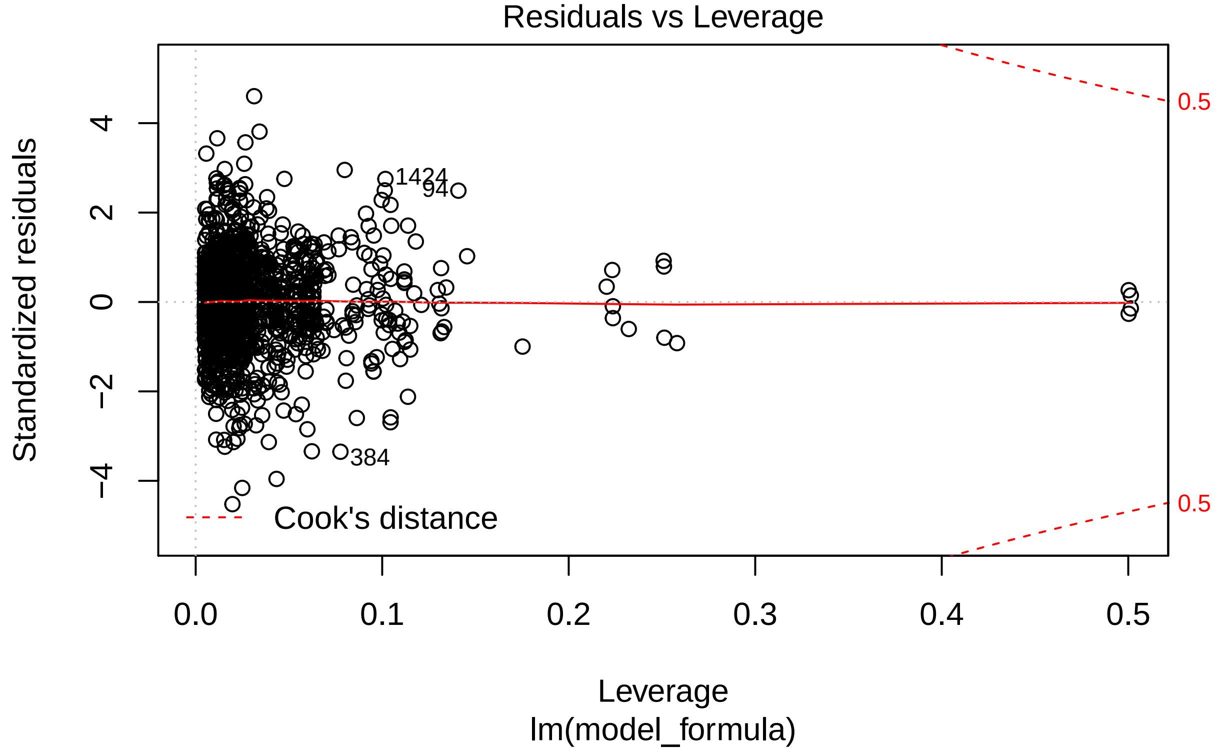
ModelName	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	devian
location	0.6668696	0.6567534	0.2340279	65.92065	0	44	71.06803	-52.13605	185.7426	77.552
inside	0.6463671	0.6296121	0.2431044	38.57748	0	67	27.46851	81.06299	440.5240	82.325
outside	0.6066970	0.5964634	0.2537499	59.28467	0	38	-50.14585	178.29170	384.4532	91.561

Based on R_2, AIC, and BIC, the best model appears to be the location model. Let's examine some diagnostic plots to make sure the assumptions of linear regression are met.


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
186
534
969
11.0
11.5
12.0
12.5
13.0
Fitted values
lm(model_formula)


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
186
534
969
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(model_formula)





Parameter Interpretation

Interpretation Confidence Intervals

Predictive Models

Restatement of problem here

Model Selection

Type of Selection Assumptions Comparing Competing Models AIC, BIC, adj R2 Interval CVPress External Cross Validation Kaggle Score

Conclusion

Appendix