Predicting House Prices with a Linear Regression Model

Kevin Thompson, Sterling Beason, & Brandon Croom

Data Science Program, Southern Methodist University, USA

Abstract

Price prediction is pivotal for real estate. Homeowners on the sell-side want to know when to sell, what to renovate, and how much profit they can expect from their efforts. Homebuyers want to know whether they are getting a fair price, where to look for homes in their budget, and the various trade-offs that accompany a purchasing decision. Real estate companies navigate both sides of real estate; hence, they too are a key stakeholder. In the first part of our analysis, we estimate the relationship between house prices, the square footage, and neighborhood location in Ames, Iowa. In the second part of our analysis, we train a linear regression model to predict house prices in Ames, Iowa.

1 Introduction

Price prediction is pivotal for real estate. Homeowners on the sell-side want to know when to sell, what to renovate, and how much profit they can expect from their efforts. Homebuyers want to know whether they are getting a fair price, where to look for homes in their budget, and the various trade-offs that accompany a purchasing decision. Real estate companies navigate both sides of real estate; hence, they too are a key stakeholder. These stakeholders utilize multiple factors related to real estate to determine the fair price for the property. These same factors can be built into a model for price prediction that assists in taking some of the guess work out of property pricing.

The analysis performed for this paper leverages a data set focused on the housing market in Ames, Iowa. This data set is from the "House Prices: Advanced Regression Techniques" Kaggle competition (Kaggle (2016)). In this competition Kaggle competitors attempt build models that best predict housing prices based on the Ames, Iowa data set. Over the course of this paper a similar approach to this specific Kaggle competition will be taken. The first part of the paper will focus on the Ames, Iowa dataset, provide the reader with detailed information about the data and any additional variables the team created for model building. Analysis of this data will then be performed in two parts. In the first part of our analysis, we estimate the relationship between house prices, the square footage, and neighborhood location in Ames, Iowa. In the second part of our analysis, we train a linear regression model to predict house prices in Ames, Iowa.

The intent of this paper is to provide the reader with an understanding of how linear regression can be applied to data and the analysis that must be undertaken to ensure a linear regression model is adequately developed.

Ramsey and Schafer (2013)

2 Ames, Iowa Data

The data used for this analysis, described in the sections below, comes from the Kaggle Competition "House Prices: Advanced Regression Techniques" (Kaggle (2016)). The data set for this competition contains housing related data for Ames, Iowa. The total data set contains 2919 observations and 80 features or variables. Although too numerous to describe here (see the Kaggle website for full descriptors (Kaggle (2016))), these 80 features relate to quantity and quality based attributes of a physical property that may interest any of the key stakeholders (prospective home buyer, home seller, real estate company/agent). For example the data provides answers to questions such as: "How many rooms in the property?", "What is the condition of the kitchen?", "What is the location of the property?", "Is there a basement?". Delving deeper, the data set breaks down into 46 categorical variables and 34 numeric variables.

The categorical variables break down into a relatively equal mix of nominal and ordinal values (23 nominal and 23 ordinal). The ordinal variables indicate a grading of various property related components such as the overall property quality, overall property condition, room specific quality and room specific conditions. The nominal values provide information on various conditions of the property such as building materials used and dwelling type. Cock (2011)

The numeric variables contain both continuous and discrete values (20 continuous and 14 discrete). The continuous variables indicate information a prospective stakeholder would like to understand such as lot size, total square footage, and specific square footage for living spaces. The discrete variables provide the prospective stakeholder with an understanding of items such as number of bedrooms, number of bathrooms, etc. Cock (2011)

In reviewing the data it was determined that a few variables could be removed for reasons noted below:

- ID - this field is a record ID field and is not informational for analysis
- Pool Quality - this field does not contain enough variation to be useful
- Miscellaneous Feature - this field contains minimial information

Data quality checks were performed across all remaining variables to address missing values, inconsistent variable names, and to ensure consistent ordering of ordinal variables across all variables of the same type. Comparison analysis was also performed to ensure the variable ordering was appropriate, as shown in figures X & X

3 Analysis Question I

3.1 Problem Statement

A local real estate company, Century 21, would like to understand the relationship between living area of a home and sale prices. Specifically, the company would like to focus only on sales in certain neighborhoods (North Ames, Edwards, and Brookside). An estimate (or estimates) of this information as well as required confidence intervals should be provided, along with verification of the model assumptions and addressing of suspicious observations. In the conclusion quantify the relationship between living area and sale price with respect to these three neighborhoods.

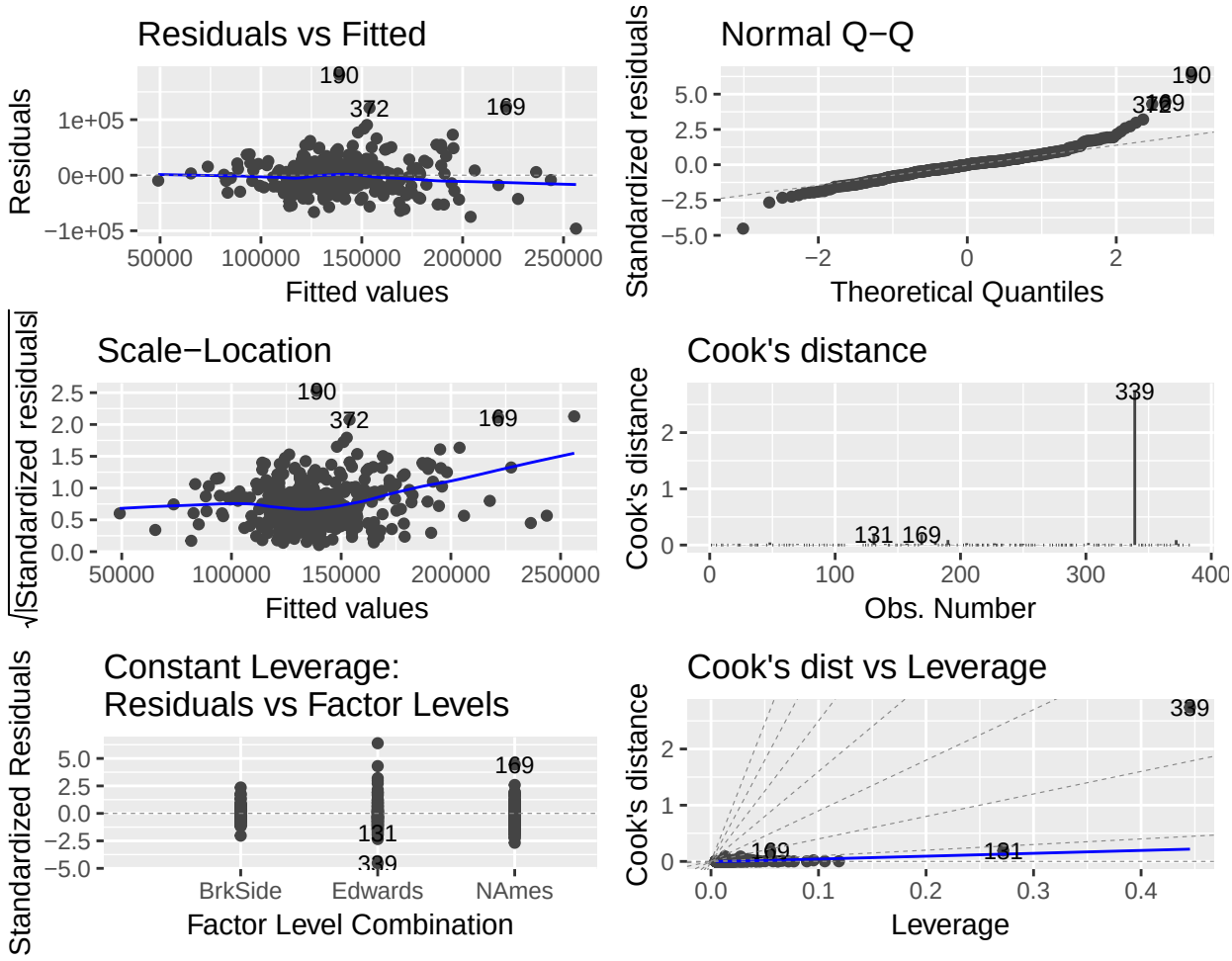
3.2 Methodology

We begin by fitting the following linear regression model:

$$Price_i = \beta_0 + \beta_1 GrLivArea_i + \beta_2 NAmes_i + \beta_3 Edwards_i + \beta_4 GrLivArea_i NAmes_i + \beta_5 GrLivArea_i Edwards_i + \epsilon_i, \tag{1}$$

where $Price_i$ denotes the price of the i-th house, $NAmes_i$ denotes whether the i-th observation is located in the North Ames neighborhood, $GrLivArea_i$ denotes the above-ground square foot living area for the i-th house, and $Edwards_i$ denotes whether the i-th observation is located in the Edwards neighborhood. The comparison neighborhood is the Brookside neighborhood. The following two interaction terms are included to capture the likely difference in the marginal contribution of living area to sale price between neighborhoods. ϵ_i denotes the error term for the i-th observation.

We examine a residual plot to verify the assumptions of the OLS estimator and the hypothesis tests for the coefficients.



The top-right qq-plot of the studentized residuals suggests moderate skew and the presence of influential observations. The bottom-right plot further clarifies the influential observation of interest to be observation

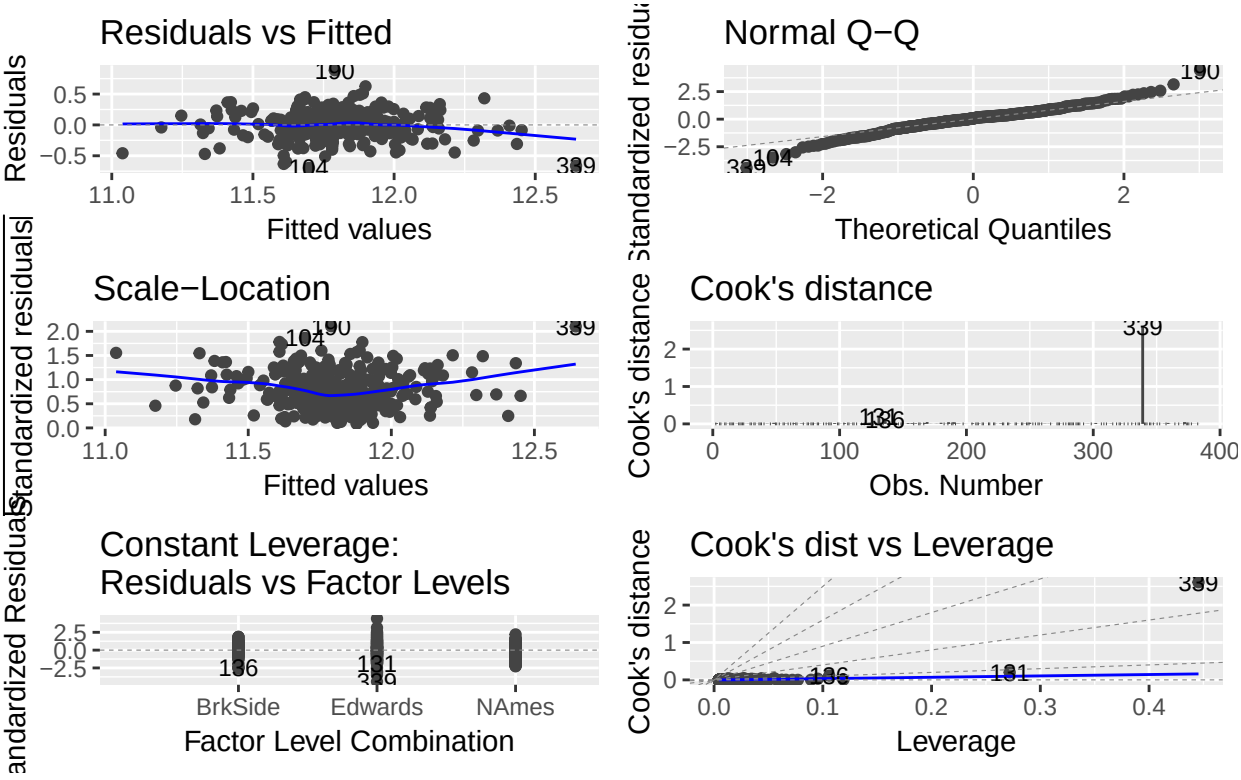
339, which has significant leverage and distance. Regressing with and without the observations yields significantly different results for the Edwards and the Edwards interaction variable, which means we cannot simply ignore it. We have no reason to believe that significant measurement error has occurred nor can we find any sufficient reason to exclude the observation, thus we turn to a log-transformation to reduce the influence of this observation. We chose the log-transformation in particular for the sake of interpretability.

Our updated model is:

$$log(Price_i) = \beta_0 + \beta_1 GrLivArea_i + \beta_2 NAmes_i + \beta_3 Edwards_i + \beta_4 GrLivArea_i NAmes_i + \beta_5 GrLivArea_i Edwards_i + \epsilon_i, \quad (2)$$

where "log" denotes the natural logarithm. We again examine the residuals to find that the relative influence of observation 339 is barely lower, though we do get similar results for our question of interest with and without the observation. Thus, in some sense, the transformation has weakened the influence of outliers.

We find no evidence of non-constant variance or non-linearity, while the number of observations gives us sufficient protection from the skewness of the residuals. We thus conclude that the log-linear model will suffice for the question of interest.




```
##
## Call:
## lm(formula = log(SalePrice) ~ GrLivArea + Neighborhood + GrLivArea *
##     Neighborhood, data = relevantData)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6963 -0.1044  0.0138  0.1107  0.8862
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.079e+01  8.702e-02 124.019  < 2e-16 ***
## GrLivArea                       7.382e-04  6.892e-05  10.712  < 2e-16 ***
## NeighborhoodNAmes               6.517e-01  9.780e-02   6.664 9.48e-11 ***
## NeighborhoodEdwards             6.303e-01  9.842e-02   6.405 4.49e-10 ***
## GrLivArea:NeighborhoodNAmes    -4.141e-04  7.620e-05  -5.435 9.88e-08 ***
## GrLivArea:NeighborhoodEdwards -5.215e-04  7.551e-05  -6.907 2.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2012 on 377 degrees of freedom
## Multiple R-squared:  0.466,  Adjusted R-squared:  0.4589
## F-statistic: 65.79 on 5 and 377 DF,  p-value: < 2.2e-16
```

We find a positive, significant association between above-ground living area square footage and the selling price of a home and that this significance is maintained between neighborhoods. We estimate that a one-hundred square foot increase in above-ground living area is associated with a 7% increase in the median selling price of a home with a 95% confidence interval of [6.02%, 8.73%]. Neither the estimate nor the significance of this variable changed with the removal of the observation, even in the base linear model. We also find that the relationship between square footage and home price changes between neighborhoods. We estimate the change in median price per additional 100 square-feet in North Ames to be 4% lower than in Brookside, while the change in median price per additional 100 square-feet in North Ames is 5% lower (after back-transforming the coefficients), with 95% confidence intervals of [-5%, -2%] and [-6%, -3%], respectively. That being said, we found that homes in North Ames and Edwards had larger median home values than Brookside. We estimate the median home value in North Ames to be 92% larger (95% confidence interval - [58%, 132%]) than Brookside and the median home value in Edwards to be 88% (95% confidence interval - [55%, 128%]) larger than in Brookside. These results cannot be generalized outside of the sample, nor can we infer a causal relationships from this observational analysis. Lastly, we must acknowledge that the presence of outliers, particularly observation 339, may have biased our estimates. Should we desire to get better estimates of the change in mean/median price for an additional 100 square-feet of living area, we will need to either get more data or add additional operational specificity to the way we define a "home".

3.3 Competing Model Comparison

3.4 Parameters

3.5 Conclusion

Ramsey and Schafer (2013)

Pearl (2009)

Ruppert and Matteson (2015)

4 Analysis Question II

4.1 Problem Statement

Build a predictive model, leveraging techniques learned in DS 6371 only, to predict sales prices of homes in all of Ames, Iowa. The goal is to produce four models: a forward selection model, a backward selection model, a stepwise selection model and a custom model. Each model should have an adjusted R^2, CV Press and Kaggle Score. In the conclusion describe which model is best at predicting future sale prices of homes in Ames, Iowa.

4.2 Assumption Checks

4.3 Competing Model Comparison

4.4 Parameters

4.5 Conclusion

Hastie et al. (2009) Trefethen and Bau (1997)

5 Appendix

References

Cock, D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression projects. reference article on Ames, Iowa data set http://www.amstat.org/publications/jse/v19n3/decock.pdf.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Kaggle (2016). House prices. data retrieved from the Kaggle website, https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Ramsey, F. and Schafer, D. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Brooks/Cole Publishing Company.

Ruppert, D. and Matteson, D. (2015). *Statistics and Data Analysis for Financial Engineering*. Springer.

Trefethen, L. and Bau, D. (1997). *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics.