# 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. Home buyers 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 data set, providing the reader with an understanding of the Exploratory Data Analysis (EDA) that occurred on the data.. 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 the 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.

## 2 Ames, Iowa Data

### 2.1 Data Description

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)

## 2.2 Exploratory Data Analysis (EDA)

Initial EDA determined that three variables could be removed for reasons noted below:

- ID (*ID*) - this field is a record ID field and is not informational for analysis
- Pool Quality (*PoolQC*) - this field does not contain enough variation to be useful
- Miscellaneous Feature (*MiscFeature*) - this field contains minimal 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.

The *LotFrontage* variable did pose a bit of a challenge for the analysis since it is a continuous numerical variable that contained NAs. *LotFrontage* is intended to represent the linear feet of street connected to the property. Initial analysis of house pricing distributions as they relate to *LotFrontage* was performed to determine impact. The distribution of house prices did not seem to change between missing and non-missing values of selling price. This provided a good indication that missingness of *LotFrontage* was unrelated to the variable of interest. To allow for *LotFrontage* to be leveraged in any models the team imputed the variable. Analysis showed that *Neighborhood*, *LotArea*, *LotShape*, and *MSSubClass* were appropriate variables to impute *LotFrontage*.

## 2.3 Data Usage

- *Training.csv* - obtained from the Kaggle site above was used to train the the linear regression models
- *Test.csv* - obtained from the Kaggle site was used to test the linear regression models

- Results - data files containing the test results of the linear regression models may be found in the GitHub repository located at: https://github.com/KThompson0308/HousePrices.git

# 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.

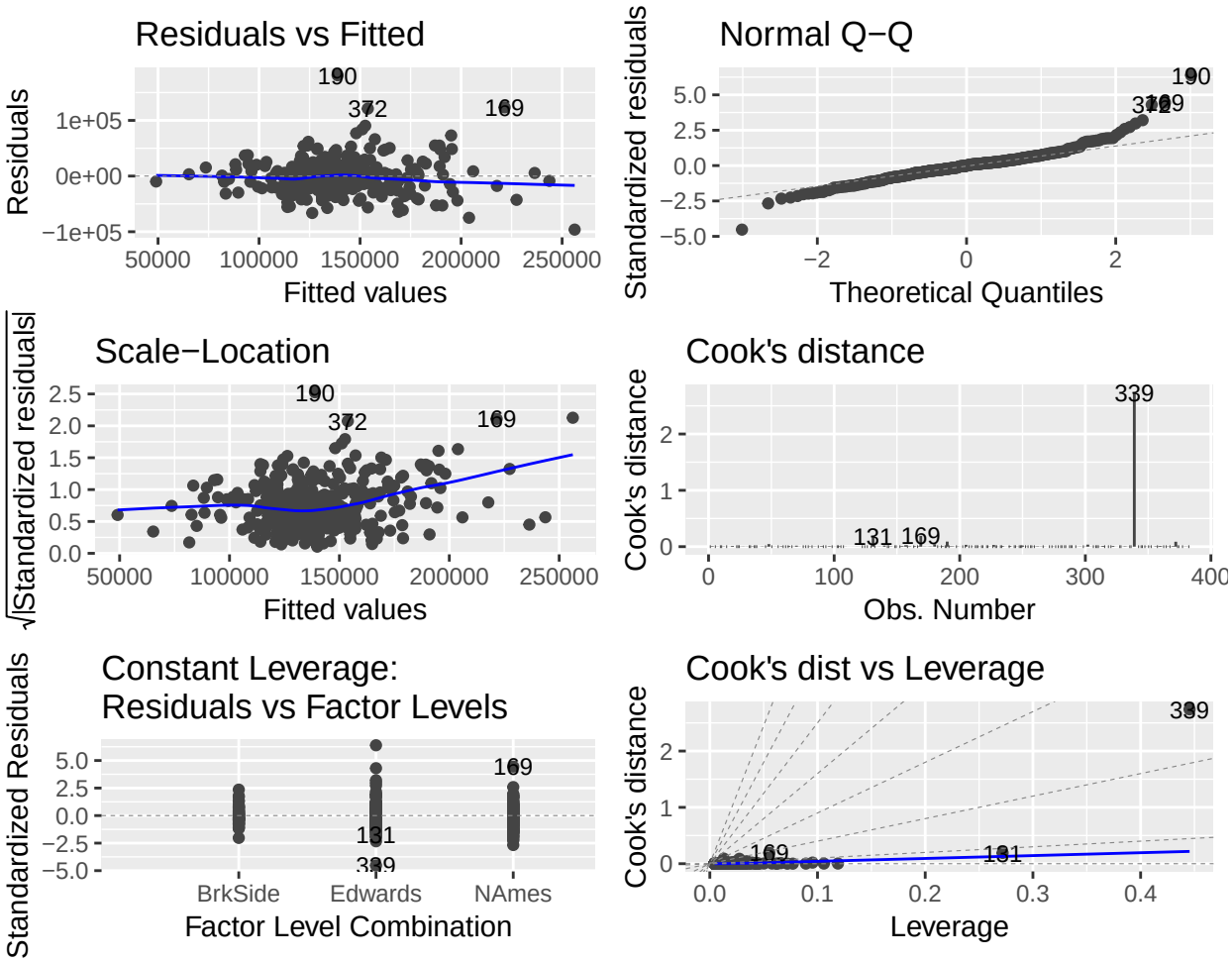
## 3.2 Methodology and Results

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. $\epsilon_i$ 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 observations of interest which have 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 chose to reduce the scope of our investigation to homes with an above ground living area of 2500 square-feet and below as the outliers of interest have extremely large values for gross living area (98th percentile and above). There are still outliers present after the removal of these observations, but their removal does not impact the results of our analysis.

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 linear model will suffice for the question of interest.

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea + Neighborhood + GrLivArea *
##     Neighborhood, data = relevantData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -96204 -14568   -310  12601 181131
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   19971.514  12351.125   1.617  0.10672
## GrLivArea                        87.163      9.782   8.911  < 2e-16 ***
## NeighborhoodNAmes             54704.888  13882.334   3.941 9.69e-05 ***
## NeighborhoodEdwards           68381.591  13969.511   4.895 1.46e-06 ***
## GrLivArea:NeighborhoodNAmes     -32.847     10.815  -3.037  0.00256 **
## GrLivArea:NeighborhoodEdwards   -57.412     10.718  -5.357 1.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28550 on 377 degrees of freedom
## Multiple R-squared:  0.4474, Adjusted R-squared:   0.44
## F-statistic: 61.04 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 $8716 increase in the mean selling price of a home with a 95% confidence interval of [6961, 10471]. We also find that the relationship between square footage and home price changes between neighborhoods. We estimate the change in mean price per additional 100 square-feet in North Ames to be $3760 lower than in Brookside,y while we did not find sufficient evidence to conclude that the change in mean price per additional 100 square-feet is different in Edwards than in Brookside. We also similarly found that homes in North Ames are estimated to have higher mean home values than Brookside. We estimate the median home value in North Ames to be $60,354 larger (95% confidence interval - [35349, 85359]) than Brookside, but we did not find sufficient evidence to conclude that mean home values are different in Edwards as opposed to Brookside. We further estimate that approximately 50% of the variation in home prices is explained by our model (Adj. $R^2$ - 0.4963) These results should not be generalized outside of the sample (homes in the relevant neighborhoods with 2500 or less square-footage of above-ground living area), nor should we infer causal relationships from this observational analysis.

# 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 Model Selection

### 4.3 Forward Selection

The model summary may be found in Table 4.1

#### 4.3.1

### 4.4 Competing Model Comparison

### 4.5 Conclusion

Model selection methods such as the forward, stepwise and backwards methods enabled model performance to be improved by reducing the number of factors to those of statistical significance. Each model selection method had various strengths as can be seen in Table X.X but the most accurate results came from the backwards model.We conclude that the backwards model was the most predictive and statistically significant model. With an Adjusted $R^2$ that is higher than the other models as well as the lowest PRESS statistic, this model is measurably more accurate than the others and therefore the one we recommend.

## 5 List of Tables

### 5.1 Analysis I

### 5.2 Analysis II

## 6 ForwardModelTable

#### 6.0.1 Forward Model Selection

#### 6.0.2 Backward Model Selection

#### 6.0.3 Stepwise Model Selection

#### 6.0.4 Custom Model Selection

## 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.

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