Predictive Model for Sales Prices of Homes (Ames, Iowa)

Using Multiple Linear Regression

August 15, 2018

Andy Ho | An Nguyen | Matthew Norton

Table of Contents

[Introduction 3](#_Toc521612843)

[Data – Descriptive Statistics 3](#_Toc521612844)

[Analysis 1 3](#_Toc521612845)

[Introduction 3](#_Toc521612846)

[Restatement of Problem 3](#_Toc521612847)

[Exploratory Statistics on Train.csv 3](#_Toc521612848)

[Selection Process 4](#_Toc521612849)

[Summary Table of Model Assumptions 4](#_Toc521612850)

[Assumptions 4](#_Toc521612851)

[Residual Plots and Influential Point 5](#_Toc521612852)

[5](#_Toc521612853)

[Model 5](#_Toc521612854)

[Fitted Model 5](#_Toc521612855)

[Parameters/Estimates 5](#_Toc521612856)

[Interpretation 5](#_Toc521612857)

[Confidence Intervals 5](#_Toc521612858)

[Conclusion 5](#_Toc521612859)

# **Introduction**

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or the presence of a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home

## Data – Descriptive Statistics

The [Ames Housing dataset](http://www.amstat.org/publications/jse/v19n3/decock.pdf) was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often-cited Boston Housing dataset.

* File Name: test.csv (461 KB) and train.csv (461 KB)
* File Origin: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>
* Data Dimensions

|  |  |  |
| --- | --- | --- |
|  | **Observations** | **Variables** |
| **Train** | 1460 | 81 |
| **Test** | 1459 | 80 |

# **Analysis 1**

## Introduction

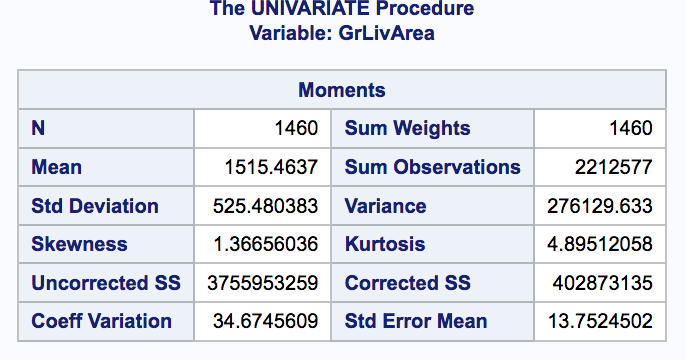
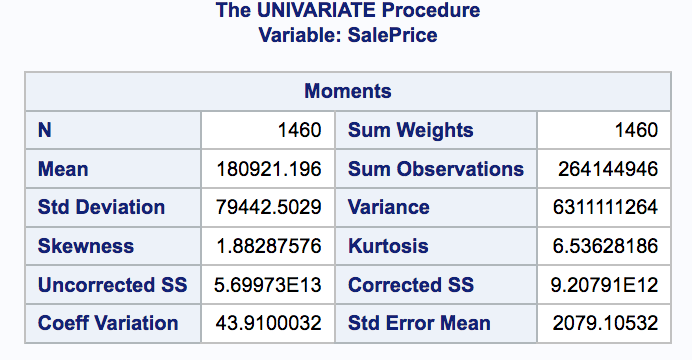
Century 21 Ames only sells home in NAmes, Edwards, and BrkSide neighborhoods and would like to get an estimate of how the SalePrice of each house is related to the square footage of the living area of the house (GrLivArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located. This portion of the report will build a model with the provided dataset and provide our conclusion that quantifies the relationship between living area and sale price with respect to the three neighborhoods.

## Restatement of Problem

Century 21 Ames would like to simply get an estimate of how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in. Using the 383 observations for the three neighborhoods, find the model that best fits the relationship between living area and sale price.

## Exploratory Statistics on Train.csv

* Descriptive Statistics: There are 1460 observations. Below are the means and standard deviations for SalePrice and GrLivArea



* Distribution of SalePrice: Evidence against normal distribution of Sale price in the three neighborhoods. Judging by the histogram, evidence of right skewness.
* Averages by neighborhoods and all three combined

|  |  |  |  |
| --- | --- | --- | --- |
| **Neighborhood** | **Number of Properties** | **Avg. Living Area** | **Avg. Sale Price** |
| NAmes | 225 | 1310.31 | $145,847.08 |
| Edwards | 100 | 1340.04 | $128,219.70 |
| BrkSide | 58 | 1203.07 | $124,834.05 |
| All Three | 383 | 1301.83 | $138,062.50 |

## Selection Process

This section will demonstrate the various models ran, results of each, and our decision on which model is best based on how it satisfies the following assumptions:

1. Normality 2. Linearity 3. Equal Variance 4. Independence 5. Outliers

## Summary Table of Model Assumptions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Vanilla Regression** | | | | | | |
| **Model** | **Normality** | **Linearity** | **Equal Variance** | **Independence** | **Outliers** | **R^2** |
| Original Data | ✔ |  |  | ✔ |  | 0.447 |
| Log-Linear | ✔ |  | ✔ | ✔ |  | 0.465 |
| Linear-Log | ✔ |  |  | ✔ |  | 0.466 |
| Log-Log | ✔ | ✔ | ✔ | ✔ |  | 0.512 |
| Original Data (Outliers Addressed) | ✔ | ✔ | ✔ | ✔ | ✔ | 0.523 |
| Log-Linear (Outliers Addressed) | ✔ | ✔ | ✔ | ✔ | ✔ | 0.525 |
| Linear-Log (Outliers Addressed) | ✔ |  | ✔ | ✔ | ✔ | 0.0498 |
| Log-Log (Outliers Addressed) | ✔ |  | ✔ | ✔ | ✔ | 0.053 |

## Assumptions

After running 8 different models, there is sufficient evidence that the vanilla regression model using the original data (three neighborhoods) while addressing the outliers is the best fit model. The following addresses the assumptions. **Refer to Addendum 5 for more charts and values**.

1. **Normality:** Judging by the histogram of the residuals, the original data (outliers addressed) there is no evidence that the residuals do not follow a normal distribution and looks slightly better than the rest.
2. **Linearity:** Judging by the scatterplots, the original data (outliers addressed) looks slightly better than the other three outliers addressed (log-linear, linear-log, log-log). The plots are randomly scattered. There is sufficient evidence for linearity.
3. **Equal Variance:** Judging by the residual scatterplots and QQplot, the original data (outliers addressed) looks better than the rest and shows sufficient evidence for equal variance.
4. **Independence:** We will assume independence, although the data gathering process was not explained.
5. **Outliers:** There are a few outliers that were addressed. Specifically, homes in the Edwards neighborhood with living area greater than 4,000 sq. ft and two homes (one in Edwards and NAmes) with sale price over $300,000.

## Residual Plots and Influential Point

## 

Residual Plots Cook’s D (middle right)

## Model

The model assigns SalePrice as the response variable and GrLivArea as the explanatory variable. Based on the scatterplots, there is strong evidence the model will have different slopes and intercepts for each significant parameter.

*PredictedSalesPrice = ß0 + ß1 \* (BrkSide) + ß2 \* (Edwards) + ß3 \* (GrLivAreaper100) + ß4 \* (BrkSide \* GrLivAreaper100) + ß5 \* (Edwards \* GrLivAreaper100)*

## Fitted Model

PredictedSalesPrice = $80,325 – $60,354 \* (BrkSide) – $43,225 \* (Edwards) + $4,956 \* (GrLivAreaper100) + $3,760 \* (BrkSide \* GrLivAreaper100) + $2,059 \* (Edwards \* GrLivAreaper100)

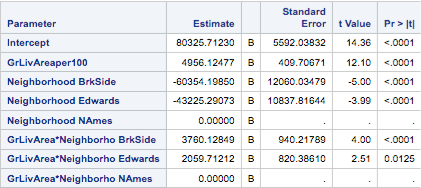
## Three Regression Equations

NAmes Neighborhood: *PredictedSalesPrice= $80,325 + $4,956 \* GrLivAreaper100*

BrkSide Neighborhood: *PredictedSalesPrice= $19,971 + $8,716 \* GrLivAreaper100*

Edwards Neighborhood: *PredictedSalesPrice= $37,100 + $7,015 \* GrLivAreaper100*

## Parameters/Estimates



### Interpretation

### Confidence Intervals

## Conclusion

After using the vanilla regression and addressing the outliers of a living area greater than 4,000 sq. ft. and a sales price of over $300,000, there is significant evidence that both the neighborhood and square footage of the living area influence the estimated sale price of the home in Ames. The intercept in this model provides an estimated sale price of $80,327 for a house with a living area of one hundred square feet in the NAmes neighborhood. For a house with a living area of one hundred square feet, the model gives us an estimated sales price decrease of $60,354 for the BrkSide neighborhood and an estimated sales price decrease of $43,225 for the Edwards neighborhood. For every increased living area of one hundred square feet the estimated sales price increases $4,956 for a home in the NAmes neighborhood, the estimated sales price increases $3,780 in the BrkSide neighborhood and the estimated sale price increases $2,059 in the Edwards neighborhood. Since this is an observational study, only associations can be made to those three neighborhoods.

# Analysis 2

## Restatement of Problem

## Model Selection

## Comparing Competing Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | .89 | 1272 | .721 |
| Backward | .78 | 1590 | .945 |
| Stepwise | .81 | 2001 | .888 |
| CUSTOM | .87 | 900 | .2345 |